# A Computational Visual Saliency Model for Perceptual Video Coding

Urooj Qureshi, Muhammad Zeeshan, Muhammad Majid and Syed Muhammad Anwar

***Abstract* – Detection of the visually salient information in videos has advantageous in various applications such as quality evaluation, compression and retrieval. Here, we present an efficient method for calculating visual saliency of videos for perceptual video coding (PVC). Our proposed algorithm is built on a block level search method. Both spatial and frequency domain saliency masks are constructed and integrated to produce the final saliency mask. In the spatial domain, the saliency mask is obtained using a variant of graph based visual saliency method. In the frequency domain, four different characteristics including motion, color, texture feature and luminance are extracted from the discrete cosine transform coefficients to generate a saliency mask. Both spatial and frequency domain masks are integrated and thresholding of saliency masks is performed. Our proposed method is evaluated on 24 test video sequences. The ground truth is prepared by manual region of interest selection sessions. The results obtained from experiments suggest that our proposed technique is computationally efficient in comparison to state of art techniques. The proposed algorithm outperforms existing techniques in terms of precision and is found to be computationally efficient.**

***Index Terms* – Visual saliency, perceptual video coding, video summarization, video quality assessment.**

## I. INTRODUCTION

HUMANS have remarkable capability in quickly identifying the most important region in a scene known as salient region. However, modeling this characteristic of human visual system (HVS) still remains a challenging task. Human vision theories suggest that we (as observer) do not emphasis on the whole image, and give priority to a small portion which in generally is visually more attractive. Visual saliency is defined as a degree to which an entity, pixel, human or constituent sticks out with respect to its neighborhood. The distinctness is often because of the variation in certain image attribute such as contrast, depth, color, texture and luminance. Visual saliency can be calculated either from a task independent and fast bottom up process or a task dependent and slower top down process [1]. Some computational models [1]–[3] presented in recent years, focus on eye fixation prediction and object detection [2], [4], image segmentation [5], object classification, image resizing [6]. These models primarily focus on predicting saliency for images. However, these models do not consider dynamic saliency which is concerned with the prediction of human focus regions while watching videos. Early work such as guided search, feature integration theory (FIT) [7], and Itti [1] suggested that visual attributes can be calculated in a side by side manner from the entire field. It either uses low- or high- level attributes or an integration of both these attributes.

Recently proposed algorithms [3], [8], [9] have described efficient results, but their main limitation is that they do not consider dynamic saliency, and hence do not work well for videos where the scene is changing quickly. The aim of this work is the recognition of salient regions within the videos. With recent development in digital communication system the human communication methods have gone highly data dependent. Video compression becomes necessary to decrease the bandwidth requirement for storage and transmission of videos. High efficiency video coding (HEVC) alone cannot attain the perceptual quality of a video coding system therefore, a perceptual model of videos is highly needed. Visual saliency pays attention to those regions that are different from the rest of the surrounding and therefore, help us in grasping the important visual information from a location. This important information in the video is classified as region of interest (ROI) and the unnecessary information is classified as non ROI. The main idea is to incorporate the features of HVS into video coding. The ROI is obtained using a visual saliency model. Therefore, the main goal is to develop a saliency based video perception model. It would have enormous utilizations in applications such as video compression [10], object recognition and video summarization. Our algorithm is based on a low level bottom up saliency method. The proposed technique is easy, rapid and outperforms other state of art methods with regard to precision and recall. Fig. 1 shows original frames, saliency maps computed using graph based visual saliency model and saliency maps of discrete cosine transform coefficients. Saliency estimation is done without any prior knowledge about the scene. A variant of graph based visual saliency model [11] along with the discrete cosine transform (DCT) model [12] is applied on a block based video dataset to detect perceptually important regions. Human labeled ground truth is used to compute precision-recall and results are compared with different state of art saliency algorithms. The video sequences employed in this work belong to versatile categories such as conversational, entertainment, sports and natural scenes etc. The ground truth is obtained by subjective experiments, where human subjects selected the ROI from the given video sequences. In this work, a saliency model for perceptual video coding (PVC) is proposed in which a spatial domain and a frequency domain map is used to generate saliency maps for videos. Our experimental results have shown that the proposed method is computationally efficient and it outperforms existing methods in terms of precision.

Urooj Qureshi, M. Zeeshan and M. Majid are with the Department of Computer Engineering and S. M. Anwar is with the Department of Software Engineering, University of Engineering and Technology, Taxila 47050, Pakistan. Email: uroojqureshi92@yahoo.com, m.majid @uettaxila.edu.pk s.anwar@uettaxila.edu.pk. Manuscript received on April 22, 2019, revised on July 19, 2019 and Aug 06, 2019 accepted on Aug 21, 2019.

 ISSN – 1999-4974

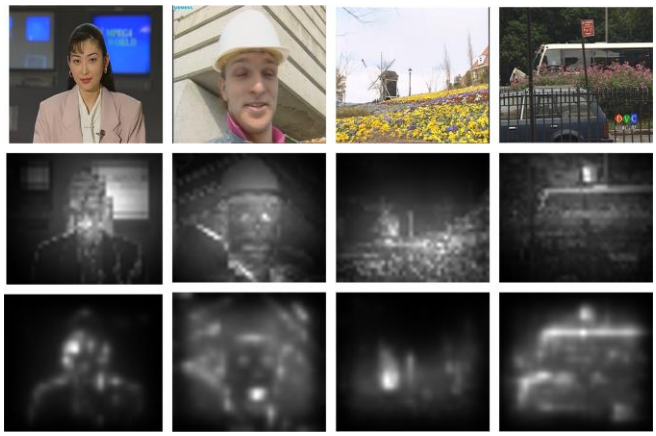

Fig. 1 Input video frames (top row), a variant of GBVS is used to compute saliency maps (middle row), saliency maps computed using DCT Coefficients (bottom row).

The rest of the paper is organized as follows. In Section II, related work is presented. The algorithm for the proposed methodology is presented in Section III. Experimental results and discussion is provided in Section IV. Concluding remarks are drawn in Section V.

## II. RELATED WORK

The human visual system is extremely fast in detecting the visual saliency however, computational modeling of this behavior is a difficult task. Saliency models aim at finding the salient region in the images or videos. The aim of saliency model is to reject the background and highlight only the important visual information. Visual attention computational models have gained significant importance in recent years. Because of these models visual communication systems have evolved. One of the main aims of object detection is to form a saliency mask by using the extracted object information. Researchers have worked on computational models that are based on prior knowledge. The second technique uses features within the images to form a saliency map. The term saliency was first defined in terms of visual attention. After that researchers have shown great interest in the detection of visual saliency. A saliency model can either be defined as a biological or a computational model, can be an integration of both these models. Itti et al. [1] presented the earliest biologically inspired system which calculated saliency maps by using center surround approach of low level characteristics such as intensity, color feature and orientation. The other methods use this model as an early inspired model. Yuan and Liang et al. [13] detected saliency maps by using object instances. Guanbin et al. [14] suggested a saliency model based on patch level approach. Frintrop et al. [15] proposed a model for the saliency map using square filters to compute the center surround difference. Integral image was also calculated to make the algorithm robust. Ma and Zhang [16] defined image saliency using local contrast analysis. The maps were further enhanced by making use of a fuzzy growth system. Achanta et al. [17] suggested a saliency detection model based on frequency domain processing. This model first calculated the average image color and then calculated the saliency by taking color difference. Liu et al. [5] combined contrast in a Gaussian image pyramid to find multi scale contrast. Goferman et al. [18] model used global consideration along with low level signal and high level attributes to detect perceptually important object and used visual organization rules to detect the object. Hu et al. [19] performed saliency estimations by using histogram thresholding of feature maps, and then heuristic measures were applied to get the final result. Cheng et al. [20] proposed histogram based contrast technique to extract saliency region and incorporated spatial relations to obtain region based contrast mask. Bruce and Tsotsos [8] defined a bottom up saliency technique build on Shannon theory and estimated self-information at each image location to acquire a saliency mask. Zhang et al. [9] also calculated saliency masks by using Shannon self-information. Murray et al. [21] described a model based on color appearances in human vision. A bank of filter was created by using the wavelet transform. Each image channel was multiplied with filters so that a pyramid of wavelet planes can be obtained. A center surround procedure and spatial pooling was used for the final mask.

Early work on dynamic saliency estimation was an extension of static techniques but recently few researchers have shown interest in video saliency recognition using spatial temporal data between two consecutive frames in a video clip. Most methods use optical flow as motion feature. Zivkovic [22] extracted video saliency using Gaussian mixture model (GMM). Li et al. [23] used temporally coherent regional paradigm for video saliency. Kim et al. [24] used textural contrast and extend this concept to spatio-temporal domain by considering temporal gradients. Fang et al. [12] predicted saliency in frequency domain where color channel, texture channel and luminance channel were obtained using DCT coefficients. Zhai and shah [25] used an algorithm for one to one point correspondence in order to describe motion contrast.

The main aim of some saliency models is to identify the salient regions in images or videos using the concept of centre bias [26]. Fu et al. [27] developed a video segmentation model that uses RGB and depth features to extract the foreground from a scene. Tao et.al [28] presented an image segmentation technique to extract important information through clustering process. Gao et al. [29] proposed a scheme based on image retrieval. The saliency maps were calculated using saliency of relevant images in the database comparing them with the overall database saliency. Sun et al. [30] saliency model was proposed for videos retrieval. It used video hash and visual features to form a saliency map. Li et al. [31] proposed an underwater scheme for image enhancement. Lei et al. [32] algorithm was based on video coding.

 ISSN – 1999-4974

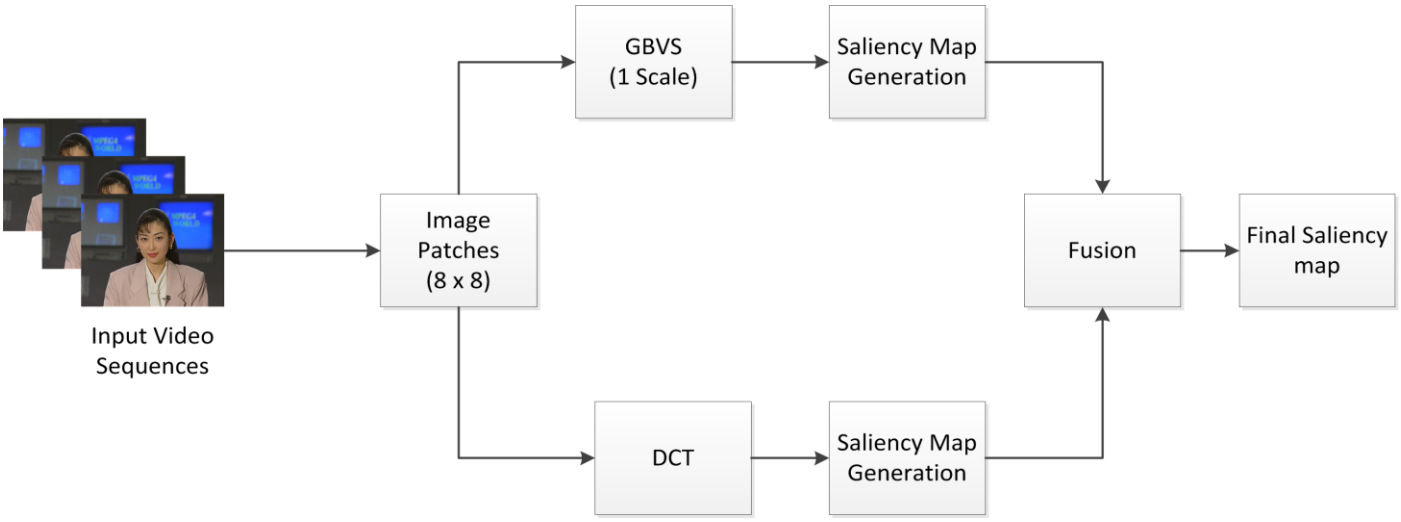


Fig. 2 Block diagram of the proposed saliency based perceptual model for perceptual video coding.

The scheme used depth maps and grayscale similarity to improve the coding performance. Recent work on saliency has been presented for image compression [31], image enhancement [33], [34], quality assessment [35], [36], action recognition [37] and thumbnail creation [38]. Cong et al. [39] presented a detailed review of different saliency detection model and addressed the issues of existing saliency models. Wand et al. [40] presented a relationship between fixation prediction and salient object segmentation. For achieving better performance and to deal with the evolving communication system models it is the need of the hour to build a specialized visual saliency model is based on perceptual video coding. In this work, a unified technique is proposed to obtain a saliency map for perceptual video coding. The proposed algorithm calculates the saliency maps of each video in 8 × 8 block level approach. The saliency maps are calculated in both spatial and frequency domain. In the spatial domain, color saliency maps are computed from each 8 × 8 block of a video frame using a variant of GBVS model. In the frequency domain, four different characteristic including motion, color, luminance and texture are extracted from each video frame in 8 × 8 block level. The color, luminance and texture feature are extracted using discrete cosine transform (DCT) coefficients. The motion feature is calculated using motion vectors. Furthermore, a fusion technique is also presented for combining both saliency maps. Due to the fusion of spatial domain saliency map and frequency domain saliency map our proposed saliency model for perceptual video coding outperforms existing state of art methods. Experiments on our proposed video coding dataset explain the benefit of our scheme with reference to modern methods.

## III. Saliency Based Perception Model

A block diagram showing the implementation detail of our saliency based perceptual model for video coding is shown in Fig. 2. Input is provided in the form of video sequences at a resolution of 352 × 288 pixels. The video sequences are then converted into frames with double precision for efficient processing. Block based approach is adopted to detect saliency, entire image is converted into 8 × 8 blocks and a saliency mask for each block is computed.

### *A. Spatial Domain Saliency Mask*

Saliency refers to the aspect of any visual stimulus that makes it stand out and gather human attention in a crowd. A variant of GBVS is used to obtain spatial domain saliency mask. Instead of multiscale feature maps, single-scale feature maps are used in our method. The size of saliency mask of each block is 4 × 4. The algorithm consists of three main steps, which are as follows:

#### *1) Feature Extraction*

The first stage is to decompose the image block into a feature map M. One feature map for intensity (I), two feature maps for color ($C_r$ and $C_b$) and two feature maps for orientation (0; 90) are calculated. Since one scale is used, there is no need to re-scale the feature maps. The feature maps are generated using biological inspired filters.

#### *2) Activation Map*

The feature maps M are used to compute the activation map A. For a given frame I, if a pixel I(a; b) is unusual in neighborhood, then the activation map will have a high value at that point. With in each feature map the difference between two regions is represented by a dissimilarity function calculated as,

$$d_1((a,b)||(p,q) = \left|\log\frac{M_1(a,b)}{M_1(p,q)}\right|, \qquad (1)$$

where $M_1(a, b)$, $M_1$ (p,q) correspond to two different region (pixels) of the feature map M. The difference between two regions is represented on logarithmic scale.

#### *3) Normalization*

In the next step, normalization and combination of activation maps is performed. The activation map is concentrated into key locations using the Markov chain and

 ISSN – 1999-4974

normalization is applied. First a connected graph ($G_A$) is established for each activation map. Each region on the activation map is treated as a node and if two nodes (a, b) and (p, q) are connected, an edge is introduced and a weight is assigned as,

$$w_1((a,b),(p,q) = d_1((a,b)||(p,q) \times F_1(a-p, b-q), \quad (2)$$

$$\mathrm{F_1(c,d)} = \exp\left(\frac{-\mathrm{c}^2 + \mathrm{d}^2}{2\sigma^2}\right) \quad (3)$$

where σ is a free parameter of the algorithm. $w_1$ is the weight of the edge, $d_1$ is the dissimilarity function and (c; d) is a node within a region on the activation map A. Each node in a $G_A$ graph is treated as a state and edge is treated as a transition probability. The weight of the outer edge for each node is normalized to 1 and an equilibrium distribution across the nodes is established. The activation map (A) is normalized using the above procedure. The activation map A is again changed into a graph ($G_n$) with nodes and weights. A Markov chain is again defined and equilibrium is established. The equilibrium distribution is the basis of the final spatial domain saliency map. The final saliency mask is computed by fusion of all feature channels, consequently all saliency blocks are combined to produce final mask for the whole frame. Gaussian kernel function is used to smooth the saliency masks.

### B. Frequency Domain Saliency Mask

Several bottom up saliency systems have been proposed to calculate the saliency by using overall information in the image [3]. Frequency domain analysis provides us a good opportunity to interpret the global information of a video frame. A saliency model based on DCT coefficients is used where the two dimensional DCT of an K×L matrix C is defined as,

$$B_{eq} = \alpha_p \alpha_q \sum_{k=0}^{K-1} \sum_{l=0}^{L-1} C_{mn} \cos\frac{\pi(2k+1)p}{2K} \cos\frac{\pi(2l+1)q}{2L}, \quad (4)$$

$$\alpha_p = \begin{cases} \frac{1}{\sqrt{K}}, & p = 0 \\ \frac{\sqrt{2}}{K}, & 1 \le p \le K-1 \end{cases} \quad (5)$$

$$\alpha_q = \begin{cases} \frac{1}{\mathrm{L}}, & \mathrm{q} = 0 \\ \frac{\sqrt{2}}{\mathrm{L}}, & 1 \le \mathrm{q} \le \mathrm{L}-1 \end{cases} \quad (6)$$

where $B_{pq}$ are DCT Coefficient of C. The first DCT coefficient (DC coefficient) reflects average energy while other coefficients (AC) reflect changes in intensity. The algorithm consists of the following steps.

#### 1) Feature Extraction

Each input video frame is converted into the lab color space to get three individual components l, a and b. Each component is divided into 8 × 8 blocks and saliency detection is performed on blocks. DCT is applied on each component to extract the feature information. DC coefficients are used to calculate the luminance and color feature for 8 × 8 block, whereas AC coefficients are used to calculate the texture information. The motion information is obtained using the predicted frames P and B, where P frame predicts motion information from a past frame, whereas B frame uses past as well as future frames for predicting motion information. The motion information is obtained from motion vector for the predicted frames.

In $YCbCr$ color space the Y channel is for luminance and two color channels i.e., $C_b$ and $C_r$. In case of videos, the DCT coefficient in 8 × 8 block is composed of 64 coefficients. The first coefficient is DC coefficient and the other 63 are AC coefficients. The luminance and color features extracted from the DCT coefficients for video are as follows,

$$L = DC_Y \ , \quad (7)$$

$$C_1 = DC_{Cr} \ , \quad (8)$$

$$C_2 = DC_{Cb} \ , \quad (9)$$

where L, $C_1$, $C_2$ are the luminance and color features of the video data. AC coefficients provide a detailed description about the frequency information. The AC coefficients from the luminance channel are used to extract texture information. First nine AC coefficients are used from the 63 coefficients. The AC coefficient is in the right corner of the discrete cosine transform block and they are equal to zero so they are ignored during quantization process. The existing studies [41] have shown that the first nine coefficients can represent most energy in each DCT block therefore, only first nine coefficients are used as follows,

$$T = AC_{01}, AC_{10}, AC_{20}, AC_{11}, \dots, AC_{30}. \quad (10)$$

The motion vectors from the video data are calculated as,

$$V = MV_p + (-1) * MV_f, \quad (11)$$

where $MV_p$ and $MV_f$ are the past and future frames respectively.

#### 2) Final Saliency Mask

The static saliency mask is obtained through the linear combination of three features including luminance, color, and texture. Normalization is applied and weight is assigned to get better results. Motion saliency mask is obtained from the motion vectors. Differentiation between motion features is computed using Euclidean distance. The final frequency domain saliency mask is acquired by integrating both static and motion saliency masks computed as follows,

$$S_f = \gamma_1 S_s t + \gamma_2 S_m + \gamma_3 S_s t S_m, \quad (12)$$


ISSN – 1999-4974

where $S_f$ is the final saliency mask, $S_s t$ is the static saliency mask and $S_m$ is the motion saliency mask. $\gamma_1, \gamma_2, \gamma_3$ are the weighted components.

### C. Spatial Frequency Integration/Fusion

The final saliency mask S is acquired by the fusion of spatial domain and frequency domain saliency masks. In this study, two usual integration methods are used and are represented as follows.

#### 1) Sum of Saliency map (SSM)

This is one of the simplest methods for combining the two maps. With $S_s$ and $S_f$ the final saliency mask S is given by,

$$SSM = S_s || S_f \,, \tag{13}$$

where $S_S$, $S_f$ is the spatial and frequency saliency mask and SSM is the final map.

#### 2) Product of Saliency map (PSM)

The fusion of two saliency maps can also be performed using the product of both maps. With $S_s$ and $S_f$ representing the spatial and frequency saliency maps, the final map PSM is given by,

$$PSM = S_s \,\&\, S_f. \tag{14}$$

#### 3) Threshold Segmentation

Saliency mask output is in the range [0; 255]. Binary masks are obtained by thresholding the saliency maps. The saliency mask is divided into sixteen threshold levels. The minimum value in case of a saliency mask is 0 and the maximum is 255. The value 0 corresponds to a black region where as the value 255 correspond to a white region. The first value 0 and the last value 255 are ignored and the other values are divided by 16 levels. To calculate the initial threshold value a small amount is added to 0 value and the same value is subtracted from the maximum 255 value. The initial threshold is then calculated as,

$$T = max - min, \tag{15}$$

The other threshold values are calculated by multiplying the initial threshold value with the level number.

## IV. RESULTS AND DISCUSSION

In this section, the outcome of our proposed perceptual video coding based saliency map method is evaluated on the video dataset. The proposed algorithm is compared with five state-of-the-art algorithms. The main goal of the generation of this large dataset is to cover videos of various motions and categories.

### A. Dataset and Ground Truth

To generate this dataset 24 video sequences are used: Bus (150 frames), City (600 frames), Tennis (150 frames), Tempete (260 frames), Soccer (600 frames), Silent (300 frames), Paris (1065 frames), News(300 frames), Mother and Daughter (300 frames), Mobile (300 frames), Miss America (150 frames), Ice (480 frames), High-way (2000 frames), Harbour (600 frames), Hall (300 frames), Football (360 frames), Foreman (300 frames), Flower (250 frames), Crew (600 frames), Container (300 frames), Coastguard (300 frames), Akiyo (300 frames), Carphone (382 frames) and Bridge-close (2001 frames). The format of the video sequences is in YUV 4:2:0. The sequences are kept at CIF (352 × 288) resolution. The video sequences cover a large number of categories like conversational, sports, entertainment and natural scenes. The videos have different types of motion information including high, medium, and low motion videos. The selection of videos was based on the fact that the sequences were frequently used in image processing and video transmission procedures. The dataset will facilitate the testing of algorithms that are based on perceptual video coding in future.

To the best of our knowledge no research has been done yet to this extent. After applying the proposed algorithm on the above-described video dataset the binary masks are then compared with the ground truth. The ground truth is obtained by subjective evaluation. 9 subjects viewed the video frames and selected the ROI using the mouse. The obtained ROI were then converted into binary masks and used as the ground truth in this research.

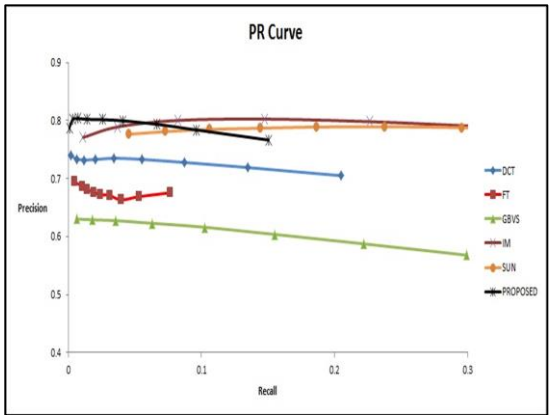


Fig. 3 Performance comparison of the proposed visual saliency model with other state-of-the-art methods in terms of precision-recall curve.

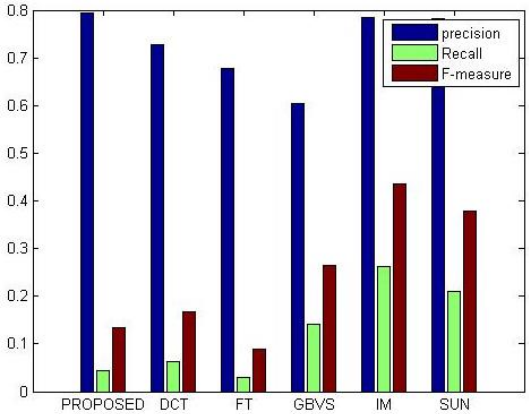


Fig. 4 Performance comparison in terms of average precision, recall and F measure of the proposed method with other visual saliency algorithms on entire video sequences dataset.


ISSN – 1999-4974

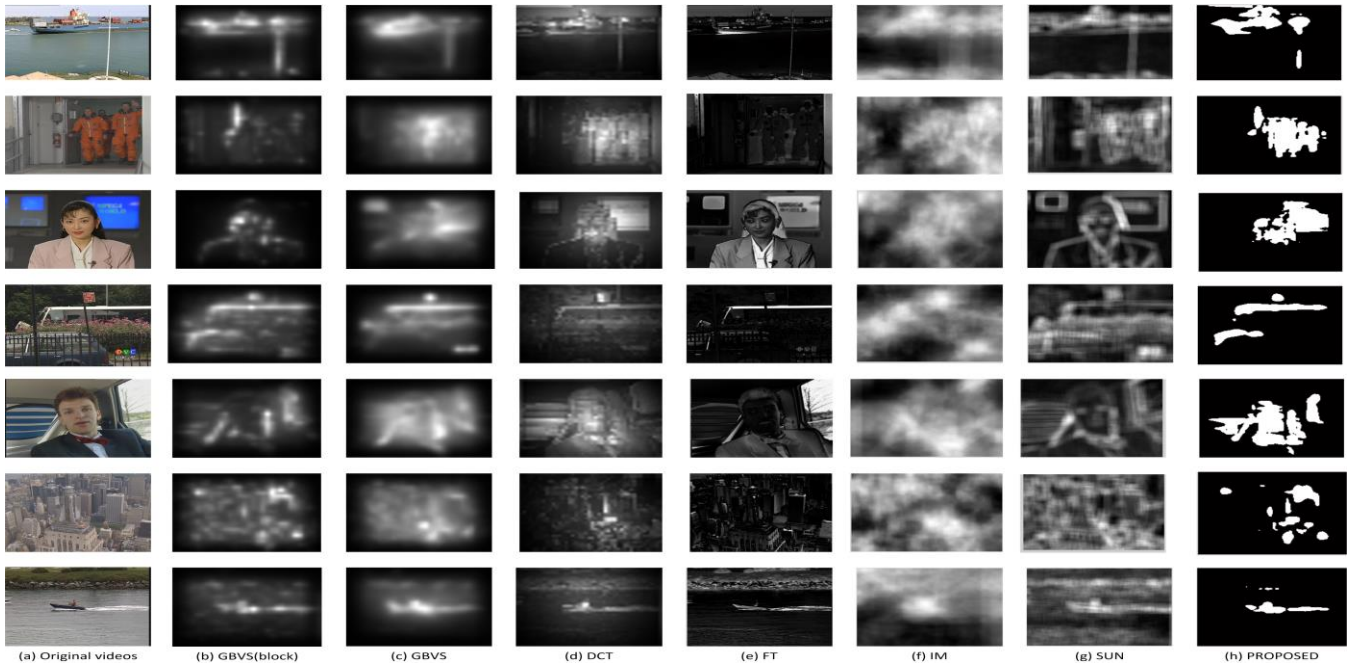


Fig. 5 Visual comparison of the saliency masks generated by the proposed method with other state-of-the-art methods.

*B. Performance Comparison*

The proposed scheme is compared with five different state-of-the-art methods. We follow Achanta [17] to select these methods. The selected methods are (IT [1], GB [11], FT [2], SR [3], SUN [9], FANG [12] and IM [21]). The methods were selected because of the variation and the recent trends. IT [1] is one of the earliest models in the saliency research area and it is a classical approach, SR [3] gives us a method in frequency domain, Fang [12] works well in compressed domain, GB [11] is the hybrid approach. SUN [9] has large number of citations. FT [2] outputs full resolution saliency maps.

After obtaining the saliency maps for all the methods the binary maps for each methods are computed using fixed thresholding method. Precision is described as the ratio of accurately allocated salient components, whereas recall is the percentage of detected salient components vs number of salient components in ground truth. Precision and recall is given by,

$$P_1 = \frac{T_p}{T_p + F_p}, \qquad (16)$$

$$R_1 = \frac{T_p}{T_p + F_n}, \qquad (17)$$

where $T_p$, $F_p$ corresponds to true positive and false positive and $F_n$ corresponds to false negative. The precision and recall curves are presented in Fig 3.

The precision and recall scores distinctly shows that the proposed method have high precision value as compared to other methods. Each point on precision recall curve represents a particular threshold level. The extreme level of the precision and recall curves is really fascinating. When the curve is near to the upper left corner it means that it has high precision and the result has higher accuracy as compared to the rest of the result. On the other hand the perfect recall score correspond to that all the salient pixels identified are relevant. The proposed method is better than other because it contains more salient pixels having value 255. Precision and recall can also be merged into a separate unit defined as F-measure, which is given by,

$$F = \frac{(1 + \beta^2) p_1 \times r_1}{\beta^2 (p_1 + r_1)}, \qquad (18)$$

where ß= 0.3.

Average F-measure along with precision and recall is shown in Fig. 4 for all the video sequences. It is evident that the proposed methods show better average F-measure when compared with other methods. Fig. 5 shows original frames, saliency maps computed for recent methods and saliency mask of the proposed model. The proposed method clearly outperforms all other state of art methods. Recall rate is not as important as precision for visual attention.

TABLE I shows the mean time used by each algorithm on Intel Pentium (R) 2.1 GHz with 2GB RAM. The author implementation is used for all the other codes. Besides producing high precision the proposed method is computationally efficient as compared to [9], [21]. The proposed method is 66 times faster as compared to [21] and it is 75 times faster as compared to [9]. The SR outputs a

 ISSN – 1999-4974

precision of 0.75. Similarly, the FT is also based on frequency domain. The FT outputs a value of 0.6766.

TABLE I
PERFORMANCE COMPARISON OF THE PROPOSED VISUAL SALIENCY MODEL WITH OTHER STATE-OF-ART VISUAL SALIENCY ALGORITHMS IN TERM OF MEAN EXECUTION TIME FOR ALL THE VIDEO SEQUENCES.

| METHOD | TIME(S) |
|---|---|
| GB | 6.0649 |
| FT | 13.07 |
| FANG | 9.996 |
| IM | 224.99 |
| SUN | 125.57 |
| PROPOSED | 33.996 |

## V. CONCLUSION

This paper suggests a new saliency model for video sequences, which identifies important regions with in the video frames. The saliency maps are computed based on spatial and frequency domain saliency maps. Both these masks are integrated leading to the final saliency mask. The spatial domain map is fast and generates results with well-defined boundaries, whereas the frequency domain map gives more information but has high computational efficiency. The basic difference between the proposed approach and existing approaches is that we consider block level analysis on videos to define a method for videos. The proposed technique is evaluated with five other state of art methods. The proposed method generates improved results in terms of precision and is computationally efficient as compared to the other methods. The proposed method gave 0.7933 precision while all the other methods have low precision values.

In future, we intend to further extend our work by encoding the saliency maps at different quantization parameters, where ROI and non-ROI are encoded at different rates. Similarly, the saliency maps can also be integrated with mode selection and sample adaptive offset filter of the video coding block. In this way optimization can be achieved. We also intend to make use of this saliency in other applications such as video summarization, video quality evaluations, and multimedia applications.

## REFERENCES


[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Transactions on pattern analysis and machine intelligence, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," Computer Vision Systems, pp. 66–75, 2008.

[3] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007, pp. 1–8.

[4] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Salient region detection by modeling distributions of color and orientation," IEEE Transactions on Multimedia, vol. 11, no. 5, pp. 892–905, 2009.

[5] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum, "Lazy snapping," in ACM Transactions on Graphics (ToG), vol. 23, no. 3. ACM, 2004, pp. 303 308.

[6] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," in ACM Transactions on graphics (TOG), vol. 26, no. 3. ACM, 2007, p. 10.

[7] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," Cognitive psychology, vol. 12, no. 1, pp. 97–136, 1980.

[8] N. D. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," Journal of vision, vol. 9, no. 3, pp. 5–5, 2009.

[9] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics, "Journal of vision, vol. 8, no. 7, pp. 32–32, 2008.

[10] C. Christopoulos, A. Skodras, and T. Ebrahimi, "The jpeg2000 still image coding system: an overview," IEEE transactions on consumer electronics, vol. 46, no. 4, pp. 1103–1127, 2000.

[11] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in Advances in neural information processing systems, 2007, pp. 545–552.

[12] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," IEEE Transactions on Image Processing, vol. 23, no. 9, pp. 3910–3921, 2014.

[13] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017, pp. 247–256.

[14] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 478–487.

[15] S. Frintrop, M. Klodt, and E. Rome, "A real-time visual attention system using integral images," in International conference on computer vision systems, vol. 25, 2007.

[16] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in Proceedings of the eleventh ACM international conference on Multimedia. ACM, 2003, pp. 374–381.

[17] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency tuned salient region detection," in Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on. IEEE, 2009, pp. 1597–1604.

[18] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 10, pp. 1915–1926, 2012.

[19] Y. Hu, X. Xie, W.-Y. Ma, L.-T. Chia, and D. Rajan, "Salient region detection using weighted feature maps based on the human visual attention model," in Pacific-Rim Conference on Multimedia.Springer, 2004, pp. 993–1000.

[20] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 3, pp. 569–582, 2015.

[21] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in Computer vision and pattern recognition (cvpr), 2011 IEEE conference on. IEEE, 2011, pp. 433–440.

[22] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol. 2. IEEE, 2004, pp. 28–31.


 ISSN – 1999-4974

[23] Y. Li, B. Sheng, L. Ma, W. Wu, and Z. Xie, "Temporally coherent video saliency using regional dynamic contrast," IEEE Transactions on Circuits and Systems for Video Technology, vol. 23, no. 12, pp. 2067–2076, 2013.
[24] W. Kim and C. Kim, "Spatiotemporal saliency detection using textural contrast and its applications," IEEE Transactions on Circuits and Systems for Video Technology, vol. 24, no. 4, pp. 646–659, 2014.
[25] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in Proceedings of the 14th ACM international conference on Multimedia. ACM, 2006, pp. 815–824.
[26] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," IEEE transactions on image processing, vol. 24, no. 12, pp. 5706–5722, 2015.
[27] H. Fu, D. Xu, and S. Lin, "Object-based multiple foreground segmentation in rgbd video," IEEE Transactions on Image Processing, vol. 26, no. 3, pp. 1418–1427, 2017.
[28] Z. Tao, H. Liu, H. Fu, and Y. Fu, "Image co segmentation via saliency guided constrained clustering with cosine similarity," in Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[29] Y. Gao, M. Shi, D. Tao, and C. Xu, "Database saliency for fast image retrieval," IEEE Transactions on Multimedia, vol. 17, no. 3, pp. 359–369, 2015.
[30] J. Sun, X. Liu, W. Wan, J. Li, D. Zhao, and H. Zhang, "Video hashing based on appearance and attention features fusion via dbn," Neurocomputing, vol. 213, pp. 84–94, 2016.
[31] C.-Y. Li, J.-C. Guo, R.-M. Cong, Y.-W. Pang, and B. Wang, "Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior," IEEE Transactions on Image Processing, vol. 25, no. 12, pp. 5664–5677, 2016.
[32] J. Lei, J. Duan, F. Wu, N. Ling, and C. Hou, "Fast mode decision based on grayscale similarity and inter-view correlation for depth map coding in 3d-hevc," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 3, pp. 706–718, 2016.
[33] J. Lei, C. Zhang, Y. Fang, Z. Gu, N. Ling, and C. Hou, "Depth sensation enhancement for multiple virtual view rendering," IEEE Transactions on Multimedia, vol. 17, no. 4, pp. 457–469, 2015.
[34] J. Lei, M. Wu, C. Zhang, F. Wu, N. Ling, and C. Hou, "Depth preserving stereo image retargeting based on pixel fusion," IEEE Transactions on Multimedia, vol. 19, no. 7, pp. 1442–1453, 2017.
[35] L. Li, Y. Zhou,W. Lin, J.Wu, X. Zhang, and B. Chen, "No-reference quality assessment of deblocked images," Neurocomputing, vol. 177, pp. 572–584, 2016.
[36] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, and W. Zhang, "Saliency-guided quality assessment of screen content images," IEEE Transactions on Multimedia, vol. 18, no. 6, pp. 1098–1110, 2016.
[37] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level cnn: saliency-aware 3-d cnn with lstm for video action recognition," IEEE Signal Processing Letters, vol. 24, no. 4, pp. 510–514, 2016.
[38] W. Wang, J. Shen, Y. Yu, and K.-L. Ma, "Stereoscopic thumbnail creation via efficient stereo saliency detection," IEEE transactions on visualization and computer graphics, vol. 23, no. 8, pp. 2014–2027, 2016.
[39] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," IEEE Transactions on Circuits and Systems for Video Technology, 2018.
[40] W. Wang, J. Shen, X. Dong, and A. Borji, "Salient object detection driven by fixation prediction," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1711–1720.
[41] X.- Y. Wang, Y.-J. Yu, and H.-Y. Yang, "An effective image retrieval scheme using color, texture and shape features," Computer Standards & Interfaces, vol. 33, no. 1, pp. 59–68, 2011

 ISSN – 1999-4974