Video saliency-detection using custom spatiotemporal fusion method

Vinay C. Warad, Ruksar Fatima
Department of Computer Science and Engineering, Khaja Bandanawaz College of Engineering, Kalaburagi, India

Article Info

Article history:

Received Jul 20, 2022
Revised Oct 15, 2022
Accepted Dec 10, 2022

Keywords:

Computing colour saliency
High-definition video compression pixel
Image saliency
Spatiotemporal diffusion
Video saliency

ABSTRACT

There have been several researches done in the field of image saliency but not as much as in video saliency. In order to increase precision and accuracy during compression, reduce coding complexity and time consumption along with memory allocation problems with our proposed solution. It is a modified high-definition video compression (HEVC) pixel based consistent spatiotemporal diffusion with temporal uniformity. It involves taking apart the video into groups of frames, computing colour saliency, integrate temporal fusion, pixel saliency fusion is conducted and then colour information guides the diffusion process for the spatiotemporal mapping with the help of permutation matrix. The proposed solution is tested on a publicly available extensive dataset with five global saliency valuation metrics and is compared with several other state-of-the-art saliency detection methods. The results display and overall best performance amongst all other candidates.

Corresponding Author:

Vinay C. Warad
Department of Computer Science and Engineering, Khaja Bandanawaz College of Engineering
Kalaburagi, Karnataka 585104, India
Email: vinay_c111@rediffmail.com

1. INTRODUCTION

The world has tried to imitate the functioning of the human eye and the brain. The marvel of the brain to distinguish among the important and non-important features of the view the eyes are seeing and take in only whatever is necessary. Various researchers have imitated this process and in today’s word, we have this in the form of conference videos, broadcasting and streaming. There have been several researches in the field of image saliency but not in video saliency. Few researches that have made a significant impact in this field. Itti’s model is one of the most [1] researched and most prominent models for image saliency. Fourier transformation is used with the help of phase spectrum and [2], [3] helps image saliency using frequency tuning. They have used the principles of inhibition of return and winner take all that is inspired from the visual nervous system [4], [5].

It is difficult for video saliency detection, as images are not still, making memory allocation and computational complexity increased. It has a video saliency detection methodology [6] that involves determining the position of an object with reference to another. They use computation of space-time-saliency map as well as computation of motion saliency map [7]-[10]. Fused static and dynamic saliency mapping [11] to obtain a space- time saliency detection model. Here dynamic texture model is employed [12] to obtain motion patterns for both stationary and dynamic scenes.

They have used fusion model but it results in low-level saliency [13]-[15]. They have used global temporal clues to forge a robust low-level saliency map [16], [17]. The disadvantage of these methodologies is that the accumulation of error is quite high and this has led to several wrong detections.

The proposed solution is a modified spatiotemporal fusion saliency detection method. It involves a spatiotemporal background to obtain high saliency values around the foreground objects. Then after ignoring the hollow effects, a series of adjustments are made to the general saliency strategies to increase efficiency of both motion and colour saliencies. The usage of cross frame super pixels and one to one spatial temporal fusion helps in overall increase in accuracy and precision during compression.

2. RELATED WORK

In this section, the works of some of the research papers that have helped in the completion of the proposed algorithm have been mentioned. This survey talks about the various video saliency methodologies along with their advantages and disadvantages [18]. Borji [19], it has also the same outline of the paper but it also includes the various aspect, which make it difficult for the algorithms to imitate the human eye-brain coordination and how to overcome them.

This paper has a notable contribution to this field of research [20]. It has a database named dynamic human fixation 1K (DHF1K) that helps in pointing out fixations that are needed during dynamic scene free viewing, then there is the attentive convolutional neural network-long short-term memory network (ACLNet) which has augmentations to the original convolutional neural network and long short-term memory (CNN-LSTM) model to enable fast end-to-end saliency learning. In this paper [21], [22] they have made some corrections in the smooth pursuits (SP) logic. It involves manual annotations of the SPs with fixation along the arithmetic points and SP salient locations by training slicing convolutional neural networks.

High-definition video compression (HEVC) system has become the new standard video compression algorithms used today. With making changes to the HEVC algorithms with the help of a spatial saliency algorithm that uses the concept of a motion vector [23], It has led to better compression and efficiency. They haves introduced a salient object segmentation that uses the combination of conditional random field (CRF) and saliency measure. It has used statistical framework and local colour contrasting, motion and illumination features [24]. Fang *et al*. [25] is also using spatiotemporal fusion with uncertainty in statistics to measure visual saliency. They have used geodesic robustness methodology to get the saliency map [26], [27]. Has been a great help to our solution formation with its super-pixel usage and adaptive colour quantization [28]-[30]. Its measurement of difference between spatial distance and histograms has helped to obtain the super-pixel saliency map. They gave us an overall idea of the various evaluation metrics to be used in this paper [31], [32]. The first section has the introduction and section 2 succeeds it with the related work [33]. Section 3 and 4 displays the proposed algorithm, its methodologies and modifications along with its final experimentation and comparison. Section 5 concludes the paper.

3. PROPOSED SYSTEM

3.1. Modeling based saliency adjustment

The robustness is obtained by combining long-term inter batch information with colour contrast computation. Background and foreground appearance models are represented by $B_M \in \mathbb{R}^{3\times bn}$ and $F_M \in \mathbb{R}^{3\times fn}$ with $bn\ and\ fn$ being their sizes respectively. The $i-th$ super pixel's RGB history in all regions is taken care of with the following equations $intra_{C_i} = \exp(\lambda - |\varphi(MC_i) - \varphi(CM_i)|)$; $\lambda = 0.5$ and $inter_{C_i} = \varphi(\frac{\min||(R_i,G_i,B_i),B_M||_2 \cdot \frac{1}{bn}\sum||(R_i,G_i,B_i),B_M||_2}{\min||(R_i,G_i,B_i),F_M||_2 \cdot \frac{1}{fn}\sum||(R_i,G_i,B_i),F_M||_2})$. Here, λ is the upper bound discrepancy degree and helps inversing the penalty between the motion and color saliencies.

3.2. Contrast-based saliency mapping

The video sequence is now divided into several short groups of frames $G_i = \{F_1, F_2, F_3, \dots, F_n\}$. Each frame F_k, where ($k$denotes the frame number) undergoes modification using simple linear iterative clustering with boundary-aware smoothing method which removes the unnecessary details. The colour and motion gradient mapping to help form the spatiotemporal gradient map with help of pixel-based computation is given by $SM_T = ||ux, uy||_2 \odot ||\nabla(F)||_2$. That is, horizontal and vertical gradient of optical flow and $\nabla(F)$ colour gradient map. We then calculate the $i-th$ super pixel's motion contrast using (1).

$$MC_i = \sum_{a_j \in \psi_i} \frac{||U_i, U_j||_2}{||a_i, a||_2}, \psi_i = \{\tau + 1 \geq ||a_i, a_j||_2 \geq \tau\} \quad (1)$$

Where l_2 norm has been used and U and a_i denote the optical flow gradient in two directions and $i-th$ super-pixel position centre respectively. ψ_i is used to denote computational contrast range and is calculated using shortest Euclidean distance between spatiotemporal map and $i-th$ superpixel.

$$\tau = \frac{r}{||\Lambda(SM_T)||_0} \sum_{\tau \in ||\tau,i|| \leq r} \left|\left|\Lambda(SM_{T_\tau})\right|\right|_0 \; ; \; l = 0.5 \min\{width, height\}, \Lambda \to down\ sampling \quad (2)$$

Colour saliency is also computed the same way as optical flow gradient, except we use the red, blue and green notations for the $i-th$ super pixel. So, the equation is $CM = \sum_{a_j \in \psi_i} \frac{\left|\left|(R_i,G_i,B_i),(R_j,G_j,B_j)\right|\right|_2}{\left|\left|a_i,a_j\right|\right|_2}$. The following equation smoothens both MC and CM as temporal and saliency value refining is done by spatial information integration.

$$CM_{k,i} \leftarrow \frac{\sum_{\tau=k-1}^{k+1} \sum_{a_{\tau,j} \in \mu\phi} \exp\left(-||c_{k,i}, c_{\tau,j}||1/\mu\right) \cdot CM_{\tau,j}}{\sum_{\tau=k-1}^{k+1} \sum_{a_{\tau,j} \in \mu\phi} \exp\left(-||c_{k,i}, c_{\tau,j}||1/\mu\right)} \quad (3)$$

Here, $c_{k,i}$ is the average of the $i-th$ super-pixel RGB colour value in $k-th$ frame while σ controls smoothing strength. The $\left|\left|a_{k,i}, a_{\tau,j}\right|\right|_2 \leq \theta$ needs to be satisfied and this is done using μ.

$$\theta = \frac{1}{m \times n} \sum_{k=1}^{n} \sum_{i=1}^{m} || \frac{1}{m} \sum_{i=1}^{m} F(SM_{T_{k,i}}), F(SM_{T_{k,i}})||_1; \; m,n = frame\ numbers \quad (4)$$

$$F(SM_{T_i}) = \begin{cases} a_i, SM_{T_i} \leq \epsilon \times \frac{1}{m} \sum_{i=1}^{m} SM_{T_i} \\ 0, otherwise \end{cases}; \; \epsilon = filter\ strenght\ control \quad (5)$$

At each batch frame level, the $q-th$ frame's smoothing rate is dynamically updated with $(1-\gamma)\theta_{s-1} + \gamma\theta_s \to \theta_s$; $\gamma = (learning\ weight, 0.2)$. Now the colour and motion saliency is integrated to get the pixel-based saliency map$LL_S = CM \odot MC$. Since this fused saliency maps increases accuracy considerably but the rate decreases, so this will be dealt with in the next section.

3.3. Accuracy boosting

Matrix M is to be considered as the input. It will be decomposed using sparse S and low level D with $\min_{D,S} \alpha\left|\left|S\right|\right|_1 + \left|\left|D\right|\right|_* \ subj = M = S + D$ where the nuclear form of D is used. With the help of robust principal component analysis (RPCA) [30] and is showcased using $S \leftarrow sign(M-D-S)[|M-D-S| - \alpha\beta]_+$ and $D \leftarrow V[\Sigma - \beta I]_+ U, (V, \Sigma, U) \leftarrow svd(Z)$. Where $svd(Z)$ denotes singular value decomposition of Lagrange multiplier and $\alpha\ and\ \beta$ represent lesser-rank and sparse threshold parameters respectively. For reduction of incorrect detections caused by the misplacement of optical flow of super pixels in the foreground's region, the given region's rough foreground is located and feature subspace of a frame k is spanned as $gI_k = \{LL_{S_{k,1}}, LL_{S_{k,2}}, \ldots .. LL_{S_{k,m}}\}$ and thus for the entire frame group we get $gB_\tau = \{gI_1, gI_2, \ldots .., gI_n\}$. This way the rough foreground is calculated as $R_{F_i} = [\sum_{k=1}^{n} LL_{S_{k,i}} - \frac{\omega}{n \times m} \sum_{k=1}^{n} \sum_{i=1}^{m} LL_{S_{k,i}}]_+$.

Here ω is reliability cotrol factor and we also get two subspaces by LL_S and RGB colour and it is given by $SB = \{cv_1, cv_2, \ldots ., cv_n\} \in \mathbb{R}^{3v \times n}$ where $cv_i = \{vec(R_{i,1}, G_{i,1}, B_{i,1}, \ldots ., R_{i,m}, G_{i,m}, B_{i,m})\}^K$ and $S_F = vec(LL_{S_1}), \ldots . vec(LL_{S_n}) \in \mathbb{R}^{v \times n}$. This helps in making a one-to-one correspondence and then pixel-based saliency mapping infusion that is dissipated on the entire group of frames. SBoverS_F causes disruptive foreground salient movements and hence with the help from [31]-[33] this issue was resolved with an alternate solution.

$$\min_{M_{cx}, S_{cx}, \vartheta, A \odot \vartheta} \left|\left|M_c\right|\right|_* + \left|\left|D_x\right|\right|_* + \left|\left|A + \vartheta\right|\right|_2 + \alpha_1\left|\left|S_c\right|\right|_1 + \alpha_2\left|\left|S_x\right|\right|; \; ||\cdot||_*$$
$$nuclear\ norm, A\ is\ position\ matrixs. t\ M_c = D_c + S_c,\ M_s = D_s + S_x,\ M_c = SB \odot \vartheta,$$
$$M_x = SF \odot \vartheta, \vartheta = \{E_1, E_2, \ldots ., E_n\}, E_i \in \{0,1\}^{m \times m}, E_i 1^K = 1. \quad (6)$$

D_c, D_xvariables represent colour and saliency mapping, ϑ is the permutation matrix while S_x, S_c represents colour feature sparse component space and saliency feature space. This entire equation set helps in correcting super-pixel correspondences.

3.4. Mathematical model

As shown in (6) generates a distributed version of convex problems $D(M_{cx}, S_{cx}, \vartheta, A \odot \vartheta) = \alpha_1||S_c||_1 + \alpha_2||E_x||_2 + \beta_1||M_c||_* + \beta_2||M_x||_* + ||A \odot \vartheta||_2 + trace\left(Z_1^K(M_c - D_c - S_c)\right) + trace\left(Z_2^K(M_x - D_x - S_x)\right) + \frac{\pi}{2}\left(||M_c - D_c - S_c||_2 + ||(M_x - D_x - S_x)||_2\right)$. Where Z_i represents Lagrangian multiplier. πdenotes steps of iterations and the optimized solution using partial derivative $S_{c,x}^{k+1} = \frac{1}{2}||S_{c,x}^k - (M_{c,x}^k - S_{c,x}^k + Z_{1,2}^k/\pi k||_2^2 + \min_{S_{c,x}^k} \alpha_{1,2} \left|\left|S_{c,x}^k\right|\right|_1 /\pi k$ and $D_{c,x}^{k+1} = \frac{1}{2}||D_{c,x}^k - (M_{c,x}^k - D_{c,x}^k + Z_{1,2}^k/\pi k||_2^2 + \min_{D_{c,x}^k} \beta_{1,2} \left|\left|D_{c,x}^k\right|\right|_* /\pi k$.

D_i is updated to become $D_{c,x}^{k+1} \leftarrow U^K + V\left[\Sigma - \frac{\beta_{1,2}}{\pi k}\right]$, where $(V, \Sigma, U) \leftarrow svd\left(M_{c,x}^k - S_{c,x}^k + \frac{Z_{1,2}^k}{\pi k}\right)$. Similarly, for S_i,$S_{c,x}^{k+1} \leftarrow sign\left(\frac{|J|}{\pi k}\right)\left[J - \frac{\alpha_{1,2}}{\pi k}\right]_+$ as $J = M_{c,x}^k - D_{c,x}^k + Z_{c,x}^k/\pi k$.

Value of E is determined are used to compute the norm cost $L \in \mathbb{R}^{m\times m}$ is calculated as $l_{i,j}^k = \left|\left|O_{k,i} - H(V_1, j)\right|\right|_2, V_1 = H(SB, k) \odot E_k$ and $l_{i,j}^k = \left|\left|O_{k,i} - H(V_2, j)\right|\right|_2, V_2 = H(SB, k) \odot E_k$. Then we use and objective matrix O to calculate the $k - th$ of R_F and the equation is $O_{k,i} = S_{c,x}(k, i) + D_{c,x}(k, i) - Z_{1,2}(k, i)/\pi k$. There is a need to change L_τ as it is hard to approximate the value of $min||A + \vartheta||_2$. $L_\tau = \{r_{1,1}^\tau + d_{1,1}^\tau, r_{1,2}^\tau + d_{1,2}^\tau, \dots, r_{m,m}^\tau + d_{m,m}^\tau\} \in \mathbb{R}^{m\times m}\ for\ k = [k - 1, k + 1]$ is hanged to L_k as shown in (7).

$$H(L_k, j) \leftarrow \sum_{\tau=k-1}^{k+1} \sum_{p_{t,v\in\xi}} H(L_\tau, v).\exp\left(-||c_{\tau,v}, c_{k,j}||\,1/\mu\right) \quad (7)$$

The global optimization is solved using the equations $SF^{k+1} \leftarrow SF^k \odot \vartheta, SB^{k+1}SB^k \odot \vartheta$ and $Z_{1,2}^{k+1} \leftarrow \pi k\left(M_{c,x}^k - D_{c,x}^k - S_{c,x}^k\right) + Z_{1,2}^k$ where $\pi_{k+1} \leftarrow \pi_k \times 1.05$. The alignment of the super pixels is now given by $gS_i = \frac{1}{n-1}\sum_{\tau=1,i\neq\tau}^n H(SF \odot \vartheta, \tau)$. To reduce the incorrect detections and alignments we introduce SF and use (8)-(10).

$$\widetilde{SF} \leftarrow SF \odot \vartheta \quad (8)$$

$$SF \leftarrow \widetilde{SF} \cdot \left(1^{m\times n} - X(S_c)\right) + \rho \cdot \overline{SF} \cdot X(S_c) \quad (9)$$

$$\rho_{i,j} = \begin{cases} 0.5, \frac{1}{n}\sum_{j=1}^n \widetilde{SF_{i,j}} < \widetilde{SF_{i,j}} \\ 2, otherwise \end{cases} \quad (10)$$

The equation for mapping for the $i - th$ video frame is given by $gS_i = \frac{H(\rho,i)-(H(\rho,i).X(S_c)}{H(\rho,i)(n-1)}\sum_{\tau=1,i\neq\tau}^n H(SF \odot \vartheta, \tau)$. There is a need to diffuse inner temporal batch x_r of the current group's frames based of degree of colour similarity. The final output is given by $gS_{i,j} = \frac{x_r \cdot y_r + \sum_{i=1}^n y_i \cdot gS_{i,j}}{y_r + \sum_{i=1}^n y_i}$; $y_r = \exp\left(-\left|\left|c_{r,j}, c_{i,j}\right|\right|_2 /\mu\right)$. Where $x_l$showcases the colour distance-based weights.

4. RESULTS, EXPERIMENTS AND DATABASE

The proposed solution has been compared with [34] as a base reference as well as by [35]'s operational block description length (OBDL) algorithm, [36]'s dynamic adaptive whitening saliency (AWS-D) algorithm, the object-to-motion convolutional neural network two layer long short-term memory (OMCNN-2CLSTM) algorithm in [36], attentive convolutional (ACL) algorithm [37], saliency-aware video compression (SAVC) algorithm from [38] and [39]. The database used is the same as the one in the base paper. It is a high-definition eye-tracking database with its open source available at GitHub *https://github.com/spzhubuaa/Video-based-Eye-Tracking-Dataset* [40]. 10 video sequences with 3 different resolutions, 1920×1080, 1280×720, and 832×480, were taken for experimentation. For evaluating the performance of all the saliency methods, we employed five global evaluation metrics, namely area under the ROC curve (AUC), Similarity (SIM), correlation coefficient (CC), normalized scanpath saliency (NSS) and Kullback-Leibler (KL).

The XU algorithm is quite similar to HEVC; hence its saliency detection is better than most algorithms but is faces problems when there are complex images as input. Other than that, our proposed solution has performed remarkably well and has the best compression efficiency and precision among all the algorithms in comparison. Table 1 shows results for saliency algorithms that are used. Figure 1 shows the saliency evaluation and comparison graph.

Table 1. The following results for saliency algorithms used: fixation maps, XU [40], base paper [34] and proposed algorithm

Parameter	BasketBall	FourPeople	RaceHorses
Fixation Maps			
XU [40]			
Base Paper [34]			
Proposed algorithm			

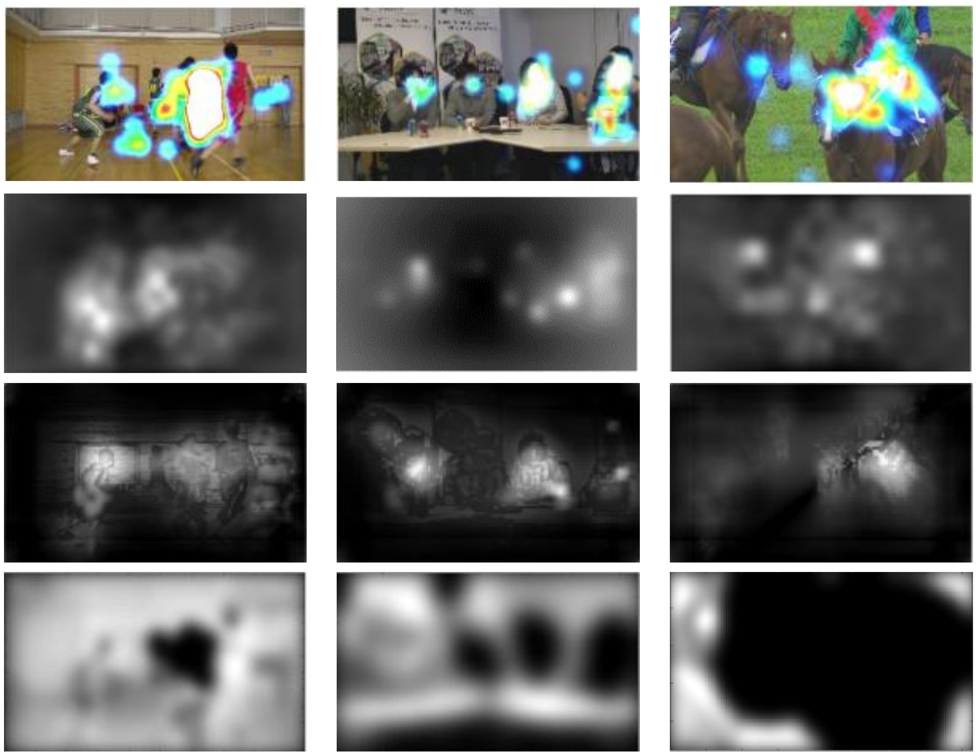

Figure 1. Saliency evaluation and comparison graph

5. CONCLUSION

This paper has proposed a solution called modified spatiotemporal fusion video saliency detection method. It involves a modified fusion calculation along with several changes to the basic HEVC code to include colour contrast computations, boost both motions, and colour values. There is also spatiotemporal of pixel-based coherency boost to increase temporal scope saliency. The proposed work is tested on the database as same as that of the base paper and is compared with other state-of-the-art methods with the help of five global evaluation metrics AUC, SIM, CC, NSS and KL. It has been concluded that the proposed algorithm of this paper has the best performance out of all the mentioned methods with better compression efficiency and precision.

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, doi: 10.1109/34.730558.

[2] C. Guo, Q. Ma, and L. Zhang, “Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform,” in *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, Jun. 2008, pp. 1–8, doi: 10.1109/CVPR.2008.4587715.

[3] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2010, pp. 1597–1604, doi: 10.1109/cvpr.2009.5206596.

[4] M. Cerf, E. P. Frady, and C. Koch, “Faces and text attract gaze independent of the task: experimental data and computer model,” *Journal of Vision*, vol. 9, no. 12, pp. 10–10, Nov. 2009, doi: 10.1167/9.12.10.

[5] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, “Predicting human gaze using low-level saliency combined with face detection,” *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, 2008.

[6] L. J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2007, pp. 1–8, doi: 10.1109/ICCV.2007.4408872.

[7] B. Scassellati, "Theory of mind for a humanoid robot," *Autonomous Robots*, vol. 12, no. 1, pp. 13–24, 2002, doi: 10.1023/A:1013298507114.

[8] S. Marat, T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Spatio-temporal saliency model to predict eye movements in video free viewing," *2008 16th European Signal Processing Conference*, Lausanne, 2008, pp. 1-5.

[9] Y. F. Ma and H. J. Zhang, "A model of motion attention for video skimming," in *IEEE International Conference on Image Processing*, 2002, vol. 1, pp. I-129-I–132, doi: 10.1109/icip.2002.1037976.

[10] S. Li and M. C. Lee, "Fast visual tracking using motion saliency in video," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2007, vol. 1, pp. I-1073-I–1076, doi: 10.1109/ICASSP.2007.366097.

[11] R. J. Peters and L. Itti, "Beyond bottom-up: incorporating task-dependent influences into a computational model of spatial attention," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2007, pp. 1–8, doi: 10.1109/CVPR.2007.383337.

[12] A. C. Schütz, D. I. Braun, and K. R. Gegenfurtner, "Object recognition during foveating eye movements," *Vision Research*, vol. 49, no. 18, pp. 2241–2253, 2009, doi: 10.1016/j.visres.2009.05.022.

[13] F. Zhou, S. B. Kang, and M. F. Cohen, "Time-mapping using space-time saliency," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 3358–3365, doi: 10.1109/CVPR.2014.429.

[14] Z. Liu, X. Zhang, S. Luo, and O. Le Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 9, pp. 1522–1540, Sep. 2014, doi: 10.1109/TCSVT.2014.2308642.

[15] Y. Li, S. Li, C. Chen, A. Hao and H. Qin, "Accurate and robust video saliency detection via self-paced diffusion," *in IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1153-1167, May 2020, doi: 10.1109/TMM.2019.2940851.

[16] Y. Fang, G. Ding, J. Li and Z. Fang, "Deep3DSaliency: deep stereoscopic video saliency detection model by 3D convolutional networks," in *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2305-2318, May 2019, doi: 10.1109/TIP.2018.2885229.

[17] C. Chen, Y. Li, S. Li, H. Qin and A. Hao, "A novel bottom-up saliency detection method for video with dynamic background," in *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 154-158, Feb. 2018, doi: 10.1109/LSP.2017.2775212.

[18] T. M. Hoang and J. Zhou, "Recent trending on learning based video compression: A survey," *Cognitive Robotics*, vol. 1, pp. 145–158, 2021, doi: 10.1016/j.cogr.2021.08.003.

[19] A. Borji, "Saliency prediction in the deep learning era: successes and limitations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 679–700, Feb. 2021, doi: 10.1109/TPAMI.2019.2935715.

[20] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 220–237, Jan. 2021, doi: 10.1109/TPAMI.2019.2924417.

[21] M. Startsev and M. Dorr, "Supersaliency: a novel pipeline for predicting smooth pursuit-based attention improves generalisability of video saliency," *IEEE Access*, vol. 8, pp. 1276–1289, 2020, doi: 10.1109/ACCESS.2019.2961835.

[22] H. Li, F. Qi, and G. Shi, "A novel spatio-temporal 3D convolutional encoder-decoder network for dynamic saliency prediction," *IEEE Access*, vol. 9, pp. 36328–36341, 2021, doi: 10.1109/ACCESS.2021.3063372.

[23] F. Guo, W. Wang, Z. Shen, J. Shen, L. Shao, and D. Tao, "Motion-aware rapid video saliency detection," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4887-4898, Dec. 2020, doi: 10.1109/TCSVT.2019.2906226..

[24] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6315 LNCS, no. PART 5, 2010, pp. 366–379, doi: 10.1007/978-3-642-15555-0_27.

[25] Y. Fang, Z. Wang, and W. Lin, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," in *Proceedings - IEEE International Conference on Multimedia and Expo*, Jul. 2013, pp. 1–6, doi: 10.1109/ICME.2013.6607572.

[26] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, vol. 07-12-June, pp. 3395–3402, doi: 10.1109/CVPR.2015.7298961.

[27] W. Wang, J. Shen, and Ling Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4185–4196, Nov. 2015, doi: 10.1109/TIP.2015.2460013.

[28] Z. Liu, L. Meur, and S. Luo, "Superpixel-based saliency detection," in *International Workshop on Image Analysis for Multimedia Interactive Services*, Jul. 2013, pp. 1–4, doi: 10.1109/WIAMIS.2013.6616119.

[29] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "what do different evaluation metrics tell us about saliency models?," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, Mar. 2019, doi: 10.1109/TPAMI.2018.2815601.

[30] J. Wright, Y. Peng, Y. Ma, A. Ganesh, and S. Rao, "Robust principal component analysis: exact recovery of corrupted low-rank matrices by convex optimization," in *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*, 2009, pp. 2080–2088.

[31] X. Zhou, C.Yang, and W.Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 597–610, 2013, doi: 10.1109/TPAMI.2012.132.

[32] Z. Zeng, T.-H. Chan, K. Jia, and D. Xu, "Finding correspondence from multiple images via sparse and low-rank decomposition," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7576 LNCS, no. PART 5, 2012, pp. 325–339, doi: 10.1007/978-3-642-33715-4_24.

[33] P. Ji, H. Li, M. Salzmann, and Y. Dai, "Robust motion segmentation with unknown correspondences," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8694 LNCS, no. PART 6, 2014, pp. 204–219, doi: 10.1007/978-3-319-10599-4_14.

[34] S. Zhu, C. Liu, and Z. Xu, "High-definition video compression system based on perception guidance of salient information of a convolutional neural network and HEVC compression domain," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 1–1, 2020, doi: 10.1109/TCSVT.2019.2911396.

[35] S. H. Khatoonabadi, N. Vasconcelos, I. V. Bajic, and Y. Shan, "How many bits does it take for a stimulus to be salient?," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, vol. 07-12-June, pp. 5501–5510, doi: 10.1109/CVPR.2015.7299189.

[36] V. Leboran, A. Garcia-Diaz, X. R. Fdez-Vidal, and X. M. Pardo, "Dynamic whitening saliency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 5, pp. 893–907, May 2017, doi: 10.1109/TPAMI.2016.2567391.

[37] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: a large-scale benchmark and a new model," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 4894–4903, doi: 10.1109/CVPR.2018.00514.

[38] H. Hadizadeh and I. V. Bajic, "Saliency-aware video compression," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 19–33, Jan. 2014, doi: 10.1109/TIP.2013.2282897.
[39] M. Xu, L. Jiang, X. Sun, Z. Ye, and Z. Wang, "Learning to detect video saliency with HEVC features," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 369–385, Jan. 2017, doi: 10.1109/TIP.2016.2628583.
[40] F. Zhang, "VED100: A video-based eye-tracking dataset on visual saliency detection," Jan 1, 2019. Distributed by Github. https://github.com/spzhubuaa/VED100-A-Video-Based-Eye-Tracking-Dataset-on-Visual-Saliency-Detection

BIOGRAPHIES OF AUTHORS

Vinay C. Warad working as assistant professor in department of computer science and engineering at Khawaja Bandanawaz College of Engineering. He has 8 years of teaching experience. His area of interest is video saliency, image retrieval. He can be contacted at email: vinaywarad999@gmail.com.

Dr. Ruksar Fatima is a professor &head of the department for computer science and engineering, Vice principal and examination in charge at Khaja Bandanawaz College of Engineering (KBNCE) Kalaburagi, Karnataka. She is the Advisory Board Member for IJESRT (International Journal of Engineering and Research Technology). She is Member of The International Association of Engineers (IAENG). She can be contacted at email: ruksarf@gmail.com.