

# Video saliency-recognition by applying custom spatio temporal fusion technique

Vinay C. Warad, Ruksar Fatima

Department of Computer Science and Engineering, KBNCE, Kalaburagi, India

## Article Info

### Article history:

Received Sep 24, 2022

Revised Jan 24, 2023

Accepted Mar 10, 2023

### Keywords:

Motion colour saliency

Pixel-based coherency

Saliency detection

Spatio-temporal

Video-saliency

## ABSTRACT

Video saliency detection is a major growing field with quite few contributions to it. The general method available today is to conduct frame wise saliency detection and this leads to several complications, including an incoherent pixel-based saliency map, making it not so useful. This paper provides a novel solution to saliency detection and mapping with its custom spatio-temporal fusion method that uses frame wise overall motion colour saliency along with pixel-based consistent spatio-temporal diffusion for its temporal uniformity. In the proposed method section, it has been discussed how the video is fragmented into groups of frames and each frame undergoes diffusion and integration in a temporary fashion for the colour saliency mapping to be computed. Then the inter group frame are used to format the pixel-based saliency fusion, after which the features, that is, fusion of pixel saliency and colour information, guide the diffusion of the spatio temporal saliency. With this, the result has been tested with 5 publicly available global saliency evaluation metrics and it comes to conclusion that the proposed algorithm performs better than several state-of-the-art saliency detection methods with increase in accuracy with a good value margin. All the results display the robustness, reliability, versatility and accuracy.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Vinay C. Warad

Department of Computer Science and Engineering, KBNCE

Kalaburagi, India

Email: vinaywarad999@gmail.com

## 1. INTRODUCTION

The human eyes have proven to be an amazing marvel of nature. The brain and eyes together make a powerful group, which can not only see 10 million distinct colors but also have a perceptive power of 50 objects per second. In general, the human eyes can focus on specific components of a picture or a video that have an importance to us. The brain in turn filters out the unnecessary bits of information this way, keeping only those that have importance. Since a video is a series of images, the amount of information to be processed increases along with the perception of dimensions.

In the technical world, this method of image and video processing has been tried to be copied or reconstructed in a different way. Looking at the available saliency models for stationary images, we have Itti's model [1], this is regarded as the most used model for stationary image. Other models such as [2], which use Fourier transformation along the lines of phase spectrum and [3] uses frequency tuning for saliency detection. The commonality among the aforementioned models is the employment of the bottom-up visual attention mechanism. For example, [3] model uses a range of frequencies in the image spectrum, which highlights the important details, to obtain the saliency map. Then, the saliency map is computed with the help of Difference of Gaussians as well as combining several band pass filters' results. Then feature conspicuity maps are constructed with the help of all low-level image features [4], [5], which is again added into the final saliency

map result with principles such as Winner Take All or Inhibition of Return, that are taken from the visual nervous system. All of these are designed for still images and not for videos. In videos, the texture feature may not be salient in a moving image while it is present in a still image. Thus, there is need for other saliency models and methods to for videos.

Videos are series of moving images called frames and its movement is in a sequence. There is a set frame rate to render a smooth motion so that the brain cannot differentiate between each image. Videos also helps in determining the position of any object with reference to another [6]. It can be inferred that video saliency is much more complex than image saliency. There have been several researches done on this field and there have been majorly two methods, one being computation of space-time saliency map and the other being the computation of motion saliency map [7]–[10]. To get spatio-metrically mapped video saliency, Peters and Itti [11] fused the ideas of static and dynamic saliency mapping to get a space-time saliency detection model. We also have [12] in which the authors proposed a dynamic texture model to get the motion patterns, even for dynamic scenes.

In general, most of the video saliency models uses bottom-up imagery as the base, which is capable of handling non-stationary videos. In addition, motion information is considered an additional saliency clue to help in detection of video saliency and to accomplish this many state-of-the-art saliency methods fuse motion saliency with colour saliency. In [13]–[15] have adopted the fusion model but the result is a low-level saliency. Almost all the latest keep temporary smoothness in the result saliency map and this helps in improving accuracy. In [16], [17] has even used global temporal clues to obtain a low-level robust saliency but these methods have error accumulation due to usage of minimization of energy framework, as it can manage the saliency consistency over a temporal scale and this leads to wrong detections. As video saliency is a lesser researched field and it has a great room for improvement and inclusion of customized models as well, with addition of limited accuracy fall while guaranteeing temporal saliency consistency.

The general method in video saliency algorithms is to use state-of-the-art image saliency detection to use as basic saliency clues but, in this paper, the chosen method is to not involve any high-level priors or constraints and only just straight low contrast saliency. The hollow effect is also avoided by integrating spatial-temporal gradient map. The temporal-level global clue is taken as the appearance modelling as this helps in guiding the motion saliency and colour saliency fusion. The proposed solution of custom spatio-temporal fusion saliency detection method, thus, is a spatial temporal gradient definition that helps in assigning high saliency values around foreground object and also not take into consideration the hollow effects. The efficiency and accuracy of the solution is boosted by using a series of adjustments in the saliency strategies, which helps in fusion of motion and colour saliencies. The temporal smoothness is first guarded by making a temporal saliency correspondence with cross-frame super pixels and then it is leveraged for further boosting the accuracy of the saliency model by employing a one-to-one spatial temporal saliency diffusion.

## 2. LITERATURE SURVEY

This section will deal with the various research papers that have been taken inspiration from to complete the proposed solution, that is, custom spatio-temporal fusion saliency detection method. As it has been previously mentioned, image saliency distinguishes the most important details in that image. There has been an exponential increase in video compression due to increase in the traffic caused by video streaming, webinars and so on. The demand of best video quality has led to the development of various video compression algorithms, which focuses on reducing the video memory space while keeping the quality in check. The usage of convolutional neural networks (CNN) has also been done in this field. A survey had been conducted on learning-based video compression methods [18] and each method's advantages and disadvantages have been discussed. Borji [19] has researched on the various deep saliency models, its benchmarks and datasets in order to help in the development of the not so researched field of video saliency. The research also notes the differences between the human level and algorithm level saliency detection accuracy and how to patch them up.

Meanwhile, in [20] there are three contributions made. Firstly, they introduced a new benchmark named dynamic human fixation 1K (DHF1K) that helps in pointing out fixations that are needed during dynamic scene free viewing. Then comes the attentive convolutional neural networks-long short-term memory (CNN-LSTM) network (ACLNet) that augments the CNN-LSTM architecture with a supervised attention mechanism to enable fast end-to-end saliency learning. This helps the CNN-LSTM to focus on learning faster end-to-end saliency methods for better temporal saliency representation across successive frames. The third contribution is that they have performed extensive experimentation on three datasets names, DHF1K, Hollywood-2, and University of Central Florida (UCF) sports dataset. The results of the experiments conducted were of great and utmost importance for the further development in the stated field.

In [21] has given a solution to reduce the error made in smooth pursuits (SPs), that is, a major eye movement type that is unique to perception of dynamic scenes. The solution employs manual annotations of SPs, and algorithmic points for fixations along with fixation of SP salient locations or saliency prediction by training slicing CNN. This solutions model is then tested on three datasets with reference to the already available methods. The result has led to greater accuracy and efficiency. There has been another model proposal that uses 3D convolutional encoder-decoder subnetworks [22] for dynamic scene saliency prediction. The result is first started with extraction of spatial and temporal features using two subnetworks and then the decoder enlarges the features in the spatial dimensions and aggregating temporal information.

High-definition video compression (HEVC) system is the new standard video compression algorithms used today. In [23] has improved the HEVC algorithms by the proposal of a spatial saliency algorithm that uses the concept of a motion vector. The motion estimation of each block during HEVC compression based on CNN is combined and adaptive dynamic fusion takes place. There is also an algorithm for a more flexible quadratic programming (QP) selection along with another algorithm to help in rate distortion optimization.

In [24] has introduced new salient object segmentation method, which combines conditional random field (CRF) and saliency measure. Being formulated by a statistical framework and local feature contrast in colour, illumination and motion information, the resultant salient map is used in CRF model using segmentation approach to define an energy minimization and recover well-defined salient objects. In [25] Also uses the combination of spatial and temporal information along with statistical uncertainty measures to detect visual saliency. The two spatial and temporal maps are merged using a spatiotemporally adaptive entropy-based uncertainty weighting approach to get one single map.

In [26] has introduced a contrast-based saliency in a pre-defined spatial temporal surrounding. Co-saliency detection using cluster algorithms is discussed in [27]. Cluster saliency is measured using spatial, corresponding and contrast and the results are obtained by fusing the single and multi-image saliency maps. There is another research [28]–[34] where computation of robust geodesic measurement is done to get the saliency mapping. In [35]–[40] has used a super pixel-based strategy and this helps in formulating our proposed custom spatio-temporal fusion saliency detection method. The image if first segmented into super pixels and undergoes adaptive colour quantization. Next, [41], [42] they measure inter-super pixel similarity based on difference between spatial distance and histograms. Then the spatial sparsity and global contrast sparsity are measured and then integrated with inter-super pixels to generate the super-pixel saliency map [43]–[47]. In [48] has helped in choosing the various evaluation metrics and methods for saliency testing. It has referenced the main papers as well and has very well explained metrics for even a layman to understand. This paper has 5 sections. The first section handles the introduction; the second section partakes in naming every reference that has helped this paper compete the solution proposed. The third section will take care of the mathematical aspect of the algorithm proposed and how each modification is put in to increase accuracy, perception and betterment in low result areas and sections 4 and 5 display the results in comparison to various saliency detection methods and the conclusion.

### 3. PROPOSED SYSTEM

The solution proposed by his paper is based on spatial temporal saliency fusion. The available state-of-the-art methods create saliency maps using frame sequence one by one. We have used the fusion of modelling and contrast-based saliencies. The two methods are briefly explained here.

#### 3.1. Modeling based saliency adjustment

To produce a robust saliency map, there is a need to combine colour contrast computation with long-term inter batch information so that the saliency of non-salient backgrounds is reduced. We shall use  $B_M \in \mathbb{R}^{3 \times bn}$  and  $F_M \in \mathbb{R}^{3 \times fn}$  to represent background model and foreground appearance model, with  $fn$  and  $bn$  being the sizes of their respective backgrounds, while their job is to take care of the  $i$ -th super pixel's RGB (Red, Green, Blue) history in all regions. Then, we follow (1) and (2).

$$\text{intra}_{C_i} = \exp(\lambda - |\varphi(MC_i) - \varphi(CM_i)|); \lambda = 0.5 \quad (1)$$

$$\text{inter}_{C_i} = \varphi\left(\frac{\min\|(R_i, G_i, B_i), B_M\|_2 \frac{1}{bn} \sum\|(R_i, G_i, B_i), B_M\|_2}{\min\|(R_i, G_i, B_i), F_M\|_2 \frac{1}{fn} \sum\|(R_i, G_i, B_i), F_M\|_2}\right) \quad (2)$$

Here,  $\lambda$  is the upper bound discrepancy degree. This helps to inverse the penalty between the motion and color saliencies.

### 3.2. Contrast- based saliency mapping

This mapping method has been inspired by [15]–[17], [27] but there have been some changes to their proposition to best suit the paper's aim. The aforementioned papers have used frame-by-frame analogy to detect saliency in them and separate the video sequence 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 [30], taken from [31] and boundary-aware smoothing method, which is inspired from, which helps in removing the computational burden and unnecessary details. The colour and motion gradient mapping from [31], [32] for obtaining the spatio-temporal gradient map and thus obtain pixel-based contrast computation given by (3).

$$SM_T = \|\|ux, uy\|\|_2 \odot \|\|\nabla(F)\|\|_2 \quad (3)$$

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 (4),

$$MC_i = \sum_{a_j \in \psi_i} \frac{\|\|U_i, U_j\|\|_2}{\|\|a_i, a_j\|\|_2}, \psi_i = \{\tau + 1 \geq \|\|a_i, a_j\|\|_2 \geq \tau\} \quad (4)$$

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.  $\psi_i$  is used to denote computational contrast range and is calculated using shortest Euclidean distance between spatio-temporal map and  $i$  – th superpixel.

$$\tau = \frac{r}{\|\|\Lambda(SM_T)\|\|_0} \sum_{\tau \in \|\|\tau, i\|\| \leq r} \|\|\Lambda(SM_{T_\tau})\|\|_0; l = 0.5 \min\{\text{width}, \text{height}\}, \Lambda \rightarrow \text{down sampling} \quad (5)$$

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.

$$CM = \sum_{a_j \in \psi_i} \frac{\|\|(R_i, G_i, B_i), (R_j, G_j, B_j)\|\|_2}{\|\|a_i, a_j\|\|_2} \quad (6)$$

$$CM_{k,i} \leftarrow \frac{\sum_{\tau=k-1}^{k+1} \sum_{a_{\tau,j} \in \mu\phi} \exp(-\|c_{k,i} \cdot c_{\tau,j}\|^{1/\mu}) \cdot CM_{\tau,j}}{\sum_{\tau=k-1}^{k+1} \sum_{a_{\tau,j} \in \mu\phi} \exp(-\|c_{k,i} \cdot c_{\tau,j}\|^{1/\mu})} \quad (7)$$

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

$$\theta = \frac{1}{m \times n} \sum_{k=1}^n \sum_{i=1}^m \|\|\frac{1}{m} \sum_{i=1}^m F(SM_{T_{k,i}}), F(SM_{T_{k,i}})\|\|_1; m, n = \text{frame numbers} \quad (8)$$

$$F(SM_{T_i}) = \begin{cases} a_i, & SM_{T_i} \leq \epsilon \times \frac{1}{m} \sum_{i=1}^m SM_{T_i}; \\ 0, & \text{otherwise} \end{cases} \quad \epsilon = \text{filter strength control} \quad (9)$$

At each batch frame level, the  $q$  – th frame's smoothing rate is dynamically updated with (10).

$$(1 - \gamma)\theta_{s-1} + \gamma\theta_s \rightarrow \theta_s; \quad \gamma = (\text{learning weight}, 0.2) \quad (10)$$

Now the colour and motion saliency is integrated to get the pixel-based saliency map.

$$LL_S = CM \odot MC \quad (11)$$

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

There is a matrix  $M$  that is the input and it needs to be decomposed, we use the help of sparse  $S$  and low level  $D$  and use this equation  $\min_{D,S} \alpha \|\|S\|\|_1 + \|\|D\|\|_* \text{ subj} = M = S + D$  where the nuclear form of  $D$  is used

and the (11) is solved with the help of robust principal component analysis (RPCA) [34] and is showcased using the two equations. Where  $\text{ssd}(Z)$  denotes singular value decomposition of Lagrange multiplier and  $\alpha$  and  $\beta$  represent lesser-rank and sparse threshold parameters respectively. Then, to reduce wrong detections due to misplaced optical flow the super-pixels contained in the given region's rough foreground is located and feature subspace of a frame  $k$  is spanned as  $\mathbf{gI}_k = \{LL_{S_{k,1}}, LL_{S_{k,2}}, \dots, LL_{S_{k,m}}\}$  and thus for the entire frame group we get  $\mathbf{gB}_\tau = \{\mathbf{gI}_1, \mathbf{gI}_2, \dots, \mathbf{gI}_n\}$ . This way the rough foreground is calculated as given in (14).

$$S \leftarrow \text{sign}(M - D - S)[|M - D - S| - \alpha\beta]_+ \quad (12)$$

$$D \leftarrow V[\Sigma - \beta I]_+ U, (V, \Sigma, U) \leftarrow \text{svd}(Z) \quad (13)$$

$$R_{F_i} = [\sum_{k=1}^n LL_{S_{k,i}} - \frac{\omega}{n \times m} \sum_{k=1}^n \sum_{i=1}^m LL_{S_{k,i}}]_+ \quad (14)$$

Here  $\omega$  is reliability control factor and we also get two subspaces for (14) spanned by  $LL_S$  and RGB colour and it is given by  $SB = \{cv_1, cv_2, \dots, cv_n\} \in \mathbb{R}^{3v \times n}$  where  $cv_i = \{\text{vec}(R_{i,1}, G_{i,1}, B_{i,1}, \dots, R_{i,m}, G_{i,m}, B_{i,m})\}^K$  and  $S_F = \text{vec}(LL_{S_1}), \dots, \text{vec}(LL_{S_n}) \in \mathbb{R}^{v \times n}$ . 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.  $S_{\text{Bover}}$  causes disruptive foreground salient movements and hence with the help from [35]–[37] this issue was resolved with an alternate solution,

$$\min_{M_{c,x}, S_{c,x}, \vartheta, A \odot \vartheta} \left( \|M_c\|_* + \|D_x\|_* + \|A + \vartheta\|_2 + \alpha_1 \|S_c\|_1 + \alpha_2 \|S_x\|_1; \|\cdot\|_* \rightarrow \right. \\ \left. \text{nuclear norm, } A \text{ is position matrix} \right) \quad (15)$$

$$\text{s.t } M_c = D_c + S_c, M_s = D_s + S_x, M_c = SB \odot \vartheta, M_x = SF \odot \vartheta,$$

$$\vartheta = \{E_1, E_2, \dots, E_n\}, E_i \in \{0,1\}^{m \times m}, E_i 1^K = 1.$$

where the estimated pixel-based mapping features over colour and saliency feature spaces are denoted by the  $D_c, D_x$  variables,  $\vartheta$  is the permutation matrix that is taken from [36], [38], while  $S_x, S_c$  represent the colour feature sparse component space and saliency feature space. This entire equation set helps in correcting super-pixel correspondences.

### 3.4. Mathematical model

In (15) is again modified using the concept of [39]. This is to generate a distributed version of convex problems and this is represented by (16). Where  $Z_i$  represents Lagrangian multiplier.  $\pi$  Denotes steps of iterations and the optimized solution using partial derivative shown in (17).

$$D(M_{c,x}, S_{c,x}, \vartheta, A \odot \vartheta) \\ = \alpha_1 \|S_c\|_1 + \alpha_2 \|E_x\|_2 + \beta_1 \|M_c\|_* + \beta_2 \|M_x\|_* + \|A \odot \vartheta\|_2 + \text{trace} \left( Z_1^K (M_c - D_c - S_c) \right) \\ + \text{trace} \left( Z_2^K (M_x - D_x - S_x) \right) + \frac{\pi}{2} \left( \|M_c - D_c - S_c\|_2 + \|(M_x - D_x - S_x)\|_2 \right). \quad (16)$$

$$S_{c,x}^{k+1} = \frac{1}{2} \|S_{c,x}^k - (M_{c,x}^k - S_{c,x}^k + Z_{1,2}^k / \pi k)\|_2^2 + \min_{S_{c,x}^k} \alpha_{1,2} \|S_{c,x}^k\|_1 / \pi k \quad (17)$$

$$D_{c,x}^{k+1} = \frac{1}{2} \|D_{c,x}^k - (M_{c,x}^k - D_{c,x}^k + Z_{1,2}^k / \pi k)\|_2^2 + \min_{D_{c,x}^k} \beta_{1,2} \|D_{c,x}^k\|_* / \pi k \quad (18)$$

$D_i$  is updated to become

$$D_{c,x}^{k+1} \leftarrow U^K + V \left[ \Sigma - \frac{\beta_{1,2}}{\pi k} \right] \quad (19)$$

$$(V, \Sigma, U) \leftarrow \text{svd} \left( M_{c,x}^k - S_{c,x}^k + \frac{Z_{1,2}^k}{\pi k} \right)$$

similarly, for  $S_i$ .

$$S_{c,x}^{k+1} \leftarrow \text{sign} \left( \frac{|||}{\pi k} \right) \left[ J - \frac{\alpha_{1,2}}{\pi k} \right]_+ \quad (20)$$

$$J = M_{c,x}^k - D_{c,x}^k + Z_{c,x}^k / \pi k$$

then the components that determine the value of E are used o compute the norm cost  $L \in \mathbb{R}^{m \times m}$

$$l_{i,j}^k = \left\| |O_{k,i} - H(V_1, j)| \right\|_2, V_1 = H(SB, k) \odot E_k \quad (21)$$

$$l_{i,j}^k = \left\| |O_{k,i} - H(V_2, j)| \right\|_2, V_2 = H(SB, k) \odot E_k$$

where, O objective column matrix which gives k – th of  $R_F$

$$O_{k,i} = S_{c,x}(k, i) + D_{c,x}(k, i) - Z_{1,2}(k, i) / \pi k \quad (22)$$

As  $\min \|A + \vartheta\|_2$  is hard to approximate, so as to calculate it there is a need to change  $L_\tau = \{r_{1,1}^\tau + d_{1,1}^\tau, r_{1,2}^\tau + d_{1,2}^\tau, \dots, r_{m,m}^\tau + d_{m,m}^\tau\} \in \mathbb{R}^{m \times m}$ ,  $k = [k - 1, k + 1]$  to  $L_k$ .

$$H(L_k, j) \leftarrow \sum_{\tau=k-1}^{k+1} \sum_{p_t, v \in \xi} H(L_\tau, v) \cdot \exp(-|c_{t,v}, c_{k,j}| / 1/\mu) \quad (23)$$

the global optimization problems is handled using the algorithm in [40] and thus modifying the (24)-(26)

$$SF^{k+1} \leftarrow SF^k \odot \vartheta, SB^{k+1} SB^k \odot \vartheta \quad (24)$$

$$Z_{1,2}^{k+1} \leftarrow \pi k (M_{c,x}^k - D_{c,x}^k - S_{c,x}^k) + Z_{1,2}^k \quad (25)$$

$$\pi_{k+1} \leftarrow \pi_k \times 1.05 \quad (26)$$

the alignment of the super pixels is now given by (27)

$$gS_i = \frac{1}{n-1} \sum_{\tau=1, i \neq \tau}^n H(SF \odot \vartheta, \tau) \quad (27)$$

SF is modified to reduce the incorrect detections and alignments

$$\widetilde{SF} \leftarrow SF \odot \vartheta \quad (28)$$

$$SF \leftarrow \widetilde{SF} \cdot (1^{m \times n} - X(S_c)) + \rho \cdot \widetilde{SF} \cdot X(S_c) \quad (29)$$

$$\rho_{i,j} = \begin{cases} 0.5, & \frac{1}{n} \sum_{j=1}^n \widetilde{SF}_{i,j} < \widetilde{SF}_{i,j} \\ 2, & \text{otherwise} \end{cases} \quad (30)$$

In (29) is a balancing equation matrix. The equation to represent the result of the saliency mapping for the i – th video frame is

$$gS_i = \frac{H(\rho, i) - (H(\rho, i) \cdot X(S_c))}{H(\rho, i)(n-1)} \sum_{\tau=1, i \neq \tau}^n H(SF \odot \vartheta, \tau) \quad (31)$$

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

$$gS_{i,j} = \frac{x_r \cdot y_r + \sum_{i=1}^n y_i gS_{i,j}}{y_r + \sum_{i=1}^n y_i}; y_r = \exp(-|c_{r,j}, c_{i,j}| / \mu) \quad (32)$$

Where  $x_i$  displays the colour distance-based weights.

## 4. RESULTS, EXPERIMENTS AND DATABASE

### 4.1. Based references and comparison

Any experiment or research is not complete without the proposed solution's actual application results. For this paper, the algorithm is compared [42] as a bare reference, followed by [43]'s operational block description length (OBDL) [43], dynamic adaptive whitening saliency (AWS-D) [44]. Object-to-motion convolutional neural network two layer long short-term memory (OMCNN-2CLSTM) [45], attentive convolutional (ACL) [46], saliency aware video compression (SAVC) algorithm from Xu *et al.* and Bylinskii *et al.* [47], [48]. These algorithms are used on the database. The database used in this paper is same as that of the reference base paper [42]. Its a high-definition eye tracking database with the open-source present in GitHub [49]. These algorithms are used widely, with a great common intermediate factor (CIF) resolution and is also based on HD non-destructive video.

### 4.2. Experiment and results

For the final comparison and evaluation, 10 sequences of videos were taken in 3 discrete resolutions of  $1920 \times 1080$ ,  $1280 \times 720$  and  $832 \times 480$ . This is shown in Table I. Then we use five evaluation metrics, namely area under the receiver operating characteristic (ROC) curve (AUC), similarity (SIM), correlation coefficient (CC), normalized scanpath saliency (NSS) and kullback-leibler (KL) and the results are shown in Table 1. Figure 1 shows the results for saliency algorithms.

The comparison among all the aforementioned algorithms (OBDL [43], AWS-D [44], OMCNN-2CLSTM [45], ACL [46], SAVC [47], XU [48], Base reference [42] and our proposed algorithm) has been numerically arranged in Table 2 and Figure 2 shows the graphical representation of the same. Five common saliency evaluation metrics have been used, the same as used in the base reference paper [42], namely area under ROC curve, similarity or histogram intersection, pearson's CC, NSS and KL divergence. Looking at the research papers, the SAVC and OBDL algorithms have been based on  $H > 264$  that has incorporated macroblock coding strategy with fixed size and is inflexible unlike the HEVC. This causes reduction in accuracy and precision. The XU algorithm is quite similar to HEVC algorithm and so it gets better results than the previous mentioned algorithms but complex pictures will lead to difficulty in saliency detection and mapping. This gives large values of KL evaluation scheme. This problem is similarly found in OMCNN-2CLSTM [45] and sACL [46] with high values of (2.82 and 3.0642 respectively). The base paper [42] has somewhat fared well than the other saliency detection methods but yet again the KL value is quite high (at 2.4921). The propose solution has done remarkably well with respect to the KL evaluation metric with an amazing result of 0.862871, that is ground truth vale is more accurate and a remarkable NSS value of nearly being unity. The other evaluation metric values are closer to each other but this paper has outperformed in several aspects, making the proposed custom spatio-temporal fusion saliency detection method a much more successful and viable saliency detection method.

Table 1. Information regarding the types of videos chosen for evaluation and comparison

Type	Resolution	Name	Frame Rate (Hz)
A	1920×1080	Basketball Drive	50
		Kimono 1	24
		Park Scene	24
		Johnny	60
B	1280×720	Kristen And Sara	60
		Four People	60
		vidyo3	60
		vidyo4	60
C	832×480	Basketball Drill	50
		Race Horses	30

Table 2. Saliency evaluation and comparison results

Method	AUC	SIM	CC	NSS	KL
OBDL [43]	0.6413	0.2982	0.2253	0.297	3.4642
AWS-D [44]	0.6635	0.3154	0.2663	0.4768	1.7144
ACL [46]	0.7673	0.3614	0.3774	0.5005	3.0642
SAVC [47]	0.5844	0.2688	0.1248	0.1889	2.0191
XU [48]	0.5881	0.305	0.2663	0.2854	1.5098
BasePaper [42]	0.7334	0.3751	0.387	0.5674	2.4921
PS-SYSTEM	0.7354	0.4644	0.44391	1.000138	0.862871

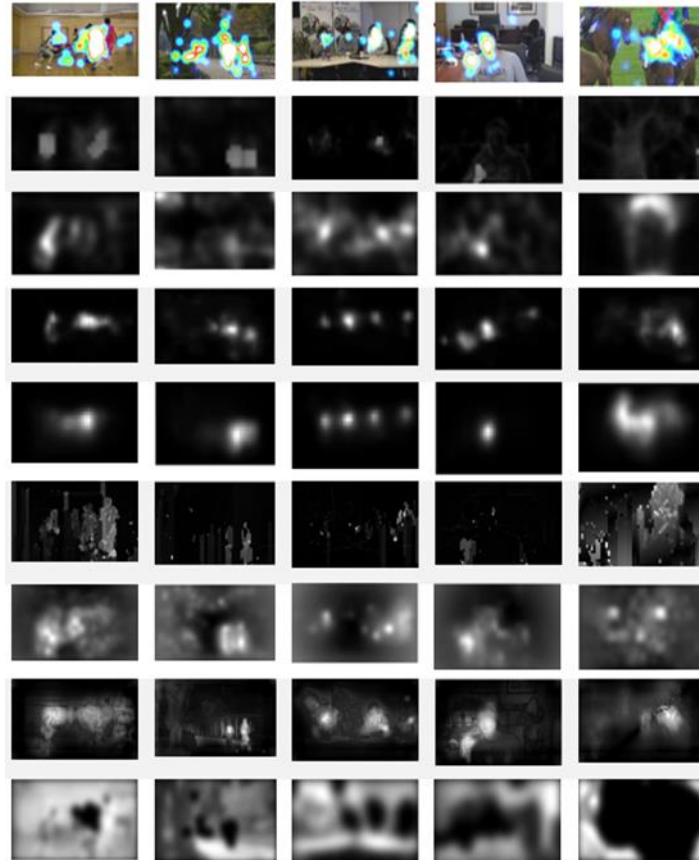


Figure 1. The results display the video frames after implementing saliency algorithm

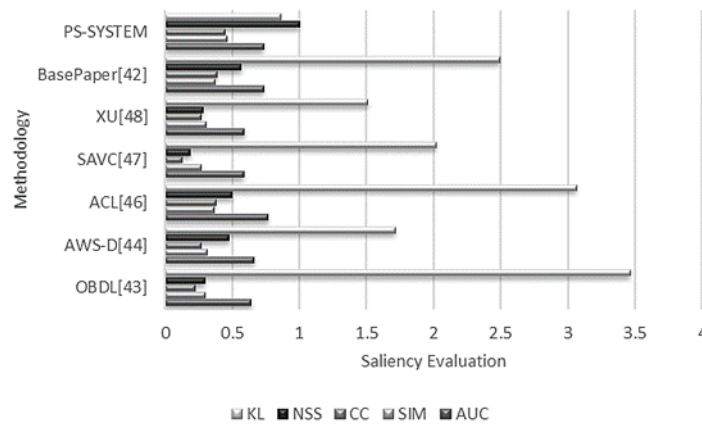


Figure 2. Saliency evaluation and comparison graph

### 5. CONCLUSION

This paper has introduced a custom spatio-temporal fusion video saliency detection method that has greater accuracy and precision in comparison to the latest available state-of-the-art saliency detection methods. There have been several changes made in simple calculations to solve the problems of colour contrast computation, modifying the fusion aspect of the saliency so as to boost both motion and colour values and also spatio-temporal of pixel-based coherency boost for temporal scope saliency exploration. The product had been tested against an extensive database provided by for comprehending its robustness and efficiency. The result has also been compared to the various state-of-the-art saliency-mapping methods and it has come to light that the proposed solution has better accuracy and precision. All these modifications have made our proposed



custom spatio-temporal fusion video saliency detection method perform much better and has given a new rise of hope in the field of video saliency. This algorithm will be helpful for those who will continue to further the research in this field of saliency detection, as there is very little research available.




## 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," *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 1–8, Jun. 2008, 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. 1–15, 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—Proceedings of the 2007 Conference*, 2008.
- [6] L. J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," *Proceedings of the IEEE International Conference on Computer Vision*, 2007, 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. Ho 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*, Aug. 2008.
- [9] Yu-Fei Ma and Hong-Jiang Zhang, "A model of motion attention for video skimming," in *Proceedings. 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 *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP '07*, 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, Sep. 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 *Proceedings of the IEEE Computer Society 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. 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, Sep. 2014, doi: 10.1109/TIP.2014.2336549.
- [16] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2015, vol. 07-12-June-2015, pp. 3395–3402, doi: 10.1109/CVPR.2015.7298961.
- [17] W. Wang, J. Shen, and L. 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.
- [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 spatiooral 3D convolutional encoder-decoder network for dynamic saliency prediction," *IEEE Access*, vol. 9, pp. 36328–36341, 2021, doi: 10.1109/ACCESS.2021.3063372.
- [23] 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. 1946–1959, 2020, doi: 10.1109/TCSVT.2019.2911396.
- [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, Springer Berlin Heidelberg, 2010, pp. 366–379.
- [25] Q. Zhang, X. Wang, S. Wang, S. Li, S. Kwong, and J. Jiang, "Learning to explore intrinsic saliency for stereoscopic video," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2019, vol. 2019-June, pp. 9741–9750, doi: 10.1109/CVPR.2019.00998.
- [26] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vision*, vol. 9, no. 12, pp. 15–15, Nov. 2009, doi: 10.1167/9.12.15.
- [27] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3766–3778, Oct. 2013, doi: 10.1109/TIP.2013.2260166.
- [28] Z. Wang, J. Li, and Z. Pan, "Cross complementary fusion network for video salient object detection," *IEEE Access*, vol. 8, pp. 201259–201270, 2020, doi: 10.1109/ACCESS.2020.3036533.
- [29] H. Bi, D. Lu, N. Li, L. Yang, and H. Guan, "Multi-level model for video saliency detection," in *2019 IEEE International Conference on Image Processing (ICIP)*, Sep. 2019, vol. 2019-Sept, pp. 4654–4658, doi: 10.1109/ICIP.2019.8803611.
- [30] E. S. L. Gastal and M. M. Oliveira, "Domain transform for edge-aware image and video processing," *ACM Transactions on Graphics*, vol. 30, no. 4, pp. 1–12, Jul. 2011, doi: 10.1145/2010324.1964964.




- [31] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012, doi: 10.1109/TPAMI.2012.120.
- [32] D. Zhang, O. Javed, and M. Shah, "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 628–635, doi: 10.1109/CVPR.2013.87.
- [33] M. Xu, P. Fu, B. Liu, and J. Li, "Multi-stream attention-aware graph convolution network for video salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 4183–4197, 2021, doi: 10.1109/TIP.2021.3070200.
- [34] 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 *NIPS'09: Proceedings of the 22nd International Conference on Neural Information Processing Systems*, 2009, pp. 2080–2088.
- [35] 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, Mar. 2013, doi: 10.1109/TPAMI.2012.132.
- [36] 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, Springer Berlin Heidelberg, 2012, pp. 325–339.
- [37] 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, Springer International Publishing, 2014, pp. 204–219.
- [38] R. Oliveira, J. Costeira, and J. Xavier, "Optimal point correspondence through the use of rank constraints," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 2, pp. 1016–1021, doi: 10.1109/CVPR.2005.264.
- [39] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010, doi: 10.1561/2200000016.
- [40] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial and Applied Mathematics*, vol. 5, no. 1, pp. 32–38, Mar. 1957, doi: 10.1137/0105003.
- [41] Z. Liu, L. Meur, and S. Luo, "Superpixel-based saliency detection," in *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, Jul. 2013, pp. 1–4, doi: 10.1109/WIAMIS.2013.6616119.
- [42] L. Wei, M. Wang, W. Liu, X. Wang, J. Sun, and X. Yin, "Multi-features fusion based on boolean map for video saliency detection," in *Chinese Control Conference, CCC*, Jul. 2019, vol. 2019-July, pp. 7589–7594, doi: 10.23919/ChiCC.2019.8865253.
- [43] 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.
- [44] F. Guo, W. Wang, Z. Shen, J. Shen, L. Shao, and D. Tao, "Motion-aware rapid video saliency detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4887–4898, Dec. 2020, doi: 10.1109/TCSVT.2019.2906226.
- [45] L. Huang, K. Song, J. Wang, M. Niu, and Y. Yan, "Multi-graph fusion and learning for RGBT image saliency detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1366–1377, Mar. 2022, doi: 10.1109/TCSVT.2021.3069812.
- [46] 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.
- [47] 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.
- [48] 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.
- [49] S. Park, E. Aksan, X. Zhang, and O. Hilliges, "Towards end-to-end video-based eye-tracking," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12357 LNCS, Springer International Publishing, 2020, pp. 747–763.

## BIOGRAPHIES OF AUTHORS



**Vinay C. Warad**    is working as assistant professor in Department of computer science and engineering at Khawaja Banda Nawaz 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.



**Ruksar Fatima**    is a professor and head of the Department for computer science and engineering, Vice principal and examination in charge at khajaBandanawaz college of engineering (KBNCE) kalaburagi Karnataka. She can be contacted at email: ruksarf@gmail.com.