Video saliency-recognition by applying custom spatio temporal fusion technique

Vinay C. Warad, Ruksar Fatima
Department of Computer Science and Engineering, KBNCE, Kalaburagi, India

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.

Keywords:
Motion colour saliency
Pixel-based coherency
Saliency detection
Spatio-temporal
Video-saliency

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 f 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 upmost 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 saliences. 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 \( B_M \in \mathbb{R}^{3 \times m} \) and \( F_M \in \mathbb{R}^{3 \times n} \) to represent background model and foreground appearance model, with \( m \) and \( n \) 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).

\[
intra_{C_1} = \exp(\lambda - |\varphi(MC_i) - \varphi(CM_i)|) \; \lambda = 0.5
\]

\[
inter_{C_1} = \varphi \left( \frac{\min(||(R_iG_i,B_i)|B_M||_2}{\max(||(R_iG_i,B_i)|F_M||_2}) \frac{1}{\frac{1}{m} \sum ||(R_iG_i,B_i)|B_M||_2} \right)
\]

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 $G_{i} = \{F_{i}, F_{i+1}, F_{i+2}, \ldots, F_{i+k}\}$, each frame $F_{i}$, where (k denotes the frame number) undergoes modification using simple linear iterative clustering [30], taken form [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).

$$\text{SM}_T = \|u_x, u_y\|_2 \otimes |\nabla(F)|_2$$  \hspace{1cm} (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),

$$MC_i = \sum_{a_i \in \psi_i} |a_i|_2, \psi_i = \{\tau + 1 \geq |a_i|_2 \geq \tau\}$$  \hspace{1cm} (4)

where $l_2$ norm has been used and $U$ and $a_i$ denote the optical flow gradient in two directions and i-th superpixel 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.

$$\tau = \frac{r}{|\text{SM}_T|^2}, \sum_{s \in ||\text{SM}_T||^2} \left| \frac{\text{SM}_T}{\text{SM}_T} \right|_2 \geq 1 = 0.5 \text{ min[width, height]}, \Lambda \rightarrow \text{down sampling}$$  \hspace{1cm} (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.

$$\text{CM} = \sum_{a_i \in \psi_i} \left| \frac{(R_{i}, G_{i}, B_{i})}{|a_i|_2} \right|_2$$  \hspace{1cm} (6)

$$\text{CM}_{k,i} = \frac{\sum_{s \in ||\text{SM}_T||^2} \sum_{c \in |\text{SM}_T|} \exp \left( -|c_{i,j}|^2 \right) \text{CM}_{i,j}^2}{\sum_{s \in ||\text{SM}_T||^2} \sum_{c \in |\text{SM}_T|} \exp \left( -|c_{i,j}|^2 \right) \text{CM}_{i,j}^2}$$  \hspace{1cm} (7)

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

$$\theta = \frac{1}{m \times n} \sum_{k=1}^{m} \sum_{i=1}^{n} |F(S_{T_{k,i}}), F(S_{T_{k,i}})|_1 \leq \epsilon \times \frac{1}{m \times n} \sum_{k=1}^{m} \sum_{i=1}^{n} S_{T_{k,i}}; \ q = \text{filter strengeth control}$$  \hspace{1cm} (8)

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

$$1 - \gamma \theta_{q-1} + \gamma \theta_q \rightarrow \theta_q; \ \gamma = \text{learning weight ,0.2}$$  \hspace{1cm} (10)

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

$$\text{LL}_S = \text{CM} \odot \text{MC}$$  \hspace{1cm} (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 $\min_{D \in S} \alpha|S|_1 + |D|_1, \text{ subj } = M = S + D$ 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 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 \( gl_k = \{ L L_{S_k1}, L L_{S_k2}, \ldots , L L_{S_{km}} \} \) and thus for the entire frame group we get \( gB_i = \{ g l_{i1}, g l_{i2}, \ldots , g l_{ia} \} \). This way the rough foreground is calculated as given in (14).

\[
S \leftarrow \text{sign}(M - D - S)[[M - D - S] - \alpha |]_+ \\
D \leftarrow V[\Sigma - \beta I]U, (V, \Sigma, U) \leftarrow \text{svd}(Z) \\
R_{Fi} = [\sum_{i=1}^{n} L L_{S_{kl}}] - \omega \pi \sum_{i=1}^{n} L L_{S_{kl}}]_+ \\
\]

(12)
(13)
(14)

Here \( \omega \) is reliability control factor and we also get two subspaces for (14) spanned by \( L L_{S} \) and RGB colour and it is given by \( S B = [c v_{1}, c v_{2}, \ldots , c v_{n}] \in \mathbb{R}^{3 \times n} \) where \( c v_{i} = \{ \text{vec}(R_{i1}, G_{i1}, B_{i1}, \ldots , R_{i m}, G_{i m}, B_{i m}) \}^{k} \) and \( S F = \text{vec}(L L_{S}), \ldots , \text{vec}(L L_{S}) \) \in \mathbb{R}^{n \times k} \). 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 over \( S_F \) causes disruptive foreground salient movements and hence with the help from [35]–[37] this issue was resolved with an alternate solution,

\[
\min_{M_c,S_c,\alpha,\theta,\theta} \|M_c\|_1 + \|D_c\|_1 + \|A + \theta\|_2 + \alpha_1\|S_c\|_1 + \alpha_2\|S_x\|_1 + \|\cdot\|_2, \rightarrow \\
\text{nuclear norm, } A \text{ is position matrix} \\
\text{s.t M}_c = D_c + S_c, M_x = D_x + S_x, M_c = S B \odot \theta, M_x = S F \odot \theta, \\
\theta = \{E_1, E_2, \ldots , E_a\}, E_i \in (0,1)^{m \times n}, E_i 1^k = 1.
\]

(15)

where the estimated pixel-based mapping features over colour and saliency feature spaces are denoted by the \( D_c, D_x \) variables, \( \theta \) is the permutation matrix that is taken from [36], [38], while \( S_C, S_x \) 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 \( Z_t \) represents Lagrangian multiplier. \( \pi \) Denotes steps of iterations and the optimized solution using partial derivative shown in (17).

\[
D(M_{c x}, S_{c x}, \alpha, A \odot \theta) \\
= \alpha_1\|S_c\|_1 + \alpha_2\|E_c\|_2 + \beta_1\|M_c\|_1 + \beta_2\|M_x\|_1 + \|A \odot \theta\|_2 + \text{trace} \left(Z_t^k(M_c - D_c - S_c)\right) \\
+ \text{trace} \left(Z_t^k(M_x - D_x - S_x)\right) + \frac{\pi}{2} \left(\|M_c - D_c - S_c\|_2 + \|M_x - D_x - S_x\|_2\right) \\
S_{c x}^{k+1} = \frac{1}{2}\|S_{c x}^k - (M_{c x}^k - S_{c x}^k + Z_{t,2}^k/\pi k)\|_2^2 + \min_{S_{c x}}\frac{\alpha_1}{2}\|S_{c x}^k\|_1/\pi k \\
D_{c x}^{k+1} = \frac{1}{2}\|D_{c x}^k - (M_{c x}^k - D_{c x}^k + Z_{t,2}^k/\pi k)\|_2^2 + \min_{D_{c x}}\frac{\beta_1}{2}\|D_{c x}^k\|_1/\pi k \\
\]

(16)
(17)
(18)

\( D_{t} \) is updated to become

\[
D_{c x}^{k+1} \leftarrow U^k + V \left[\Sigma - \beta/\pi k\right] \\
(V, \Sigma, U) \leftarrow \text{svd} \left(M_{c x}^k - S_{c x}^k + Z_{t,2}^k/\pi k\right)
\]

(19)
similarly, for $S_l$:

$$S_x^{k+1} \leftarrow \text{sign} \left( \frac{\|J\|}{\pi k} \right) \left[ J - \frac{\alpha_{1,2}}{\pi k} \right]_+$$  \hspace{1cm} (20)

$$J = \mathbf{M}^k_{cx} - \mathbf{D}^k_{cx} + \mathbf{Z}_{cx}^k/\pi k$$

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

$$I_{l_1}^k = \left\| \mathbf{O}_{k,l} - H(V_{1,l}) \right\|_2, V_1 = H(SB, k) \odot E_k$$

$$I_{l_2}^k = \left\| \mathbf{O}_{k,l} - H(V_{2,l}) \right\|_2, V_2 = H(SB, k) \odot E_k$$

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

$$O_{k,l} = S_{cx}(k,i) + D_{cx}(k,i) - Z_{1,2}(k,i)/\pi k$$  \hspace{1cm} (22)

As min$||A + \theta||_2$ is hard to approximate, so as to calculate it there is a need to change

$$L_t = \{ r_{1,1}^1 + d_{t,1} + r_{1,2}^1 + d_{t,2}, \ldots, r_{m,m}^n + d_{t,m} \} \in \mathbb{R}^{m \times m}, k = [k - 1, k + 1] \text{ to } L_k.$$  

$$H(L_{k,t}) \leftarrow \sum_{r=1}^{k+1} \sum_{c_{t,v} \in L} H(L_{t,v}), \text{ exp } (-||c_{t,v}, c_{k,t}||_1/\mu)$$  \hspace{1cm} (23)

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

$$SF^{k+1} \leftarrow SF^k \odot \emptyset, \ \text{SB}^{k+1} \odot \emptyset$$  \hspace{1cm} (24)

$$Z_{1,2}^{k+1} \leftarrow \pi k(M^k_{cx} - D^k_{cx} - S^k_{cx}) + Z_{1,2}^{k}$$  \hspace{1cm} (25)

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

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

$$g_{S_l} = \frac{1}{n-1} \sum_{r=1}^{n} \sum_{c_{t,v} \in L} H(SF \emptyset, \tau)$$  \hspace{1cm} (27)

$SF$ is modified to reduce the incorrect detections and alignments

$$\overline{SF} \leftarrow SF \odot \emptyset$$  \hspace{1cm} (28)

$$SF \leftarrow \overline{SF} \cdot (1^{m \times n} \times X(S_c)) + \rho \cdot \overline{SF} \times X(S_c)$$  \hspace{1cm} (29)

$$\rho_{ij} = \begin{cases} 0.5, & \sum_{t=1}^{n} SF_{tij} < \overline{SF}_{tij} \\ 2, & \text{otherwise} \end{cases}$$  \hspace{1cm} (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

$$g_{S_l} = \frac{H(p, i) - (H(p, i) \times X_S)}{H(p, i) \times (n - 1)} \sum_{t=1}^{n} \sum_{c_{t,v} \in L} H(SF \emptyset, \tau)$$  \hspace{1cm} (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

$$g_{S_{l,ij}} = \frac{x_r y_r + \sum_{t=1}^{n} y_i g_{S_l}^{t,ij}}{y_r + \sum_{t=1}^{n} y_i}; \ y_r = \text{ exp } (-\|c_{r,i}, c_{ij}\|_2/\mu)$$  \hspace{1cm} (32)

Where $x_r$ 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 × 1080, 1280 × 720 and 832 × 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 I. 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 value 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>
<thead>
<tr>
<th>Table 1. Information regarding the types of videos chosen for evaluation and comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type</strong></td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. Saliency evaluation and comparison results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Method</strong></td>
</tr>
<tr>
<td>OBDL [43]</td>
</tr>
<tr>
<td>AWS-D [44]</td>
</tr>
<tr>
<td>ACL [46]</td>
</tr>
<tr>
<td>SAVC [47]</td>
</tr>
<tr>
<td>Xu [48]</td>
</tr>
<tr>
<td>BasePaper [42]</td>
</tr>
<tr>
<td>PS-SYSTEM</td>
</tr>
</tbody>
</table>

Int J Artif Intell, Vol. 13, No. 1, March 2024: 82-91
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


**BIographies of Authors**

Vinay C. Warad 🇮🇳 🇱🇷 🇺🇸 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.

Rukasar Fatima 🇮🇳 🇵🇰 🇱🇷 is a professor and head of the Department for computer science and engineering. Vice principal and examination in charge at khabjabanawaz college of engineering (KBNCE) kalaburagi Karnataka. She can be contacted at email: ruksarfat@gmail.com.