Efficient fusion of spatio-temporal saliency for frame wise saliency identification

Sharada P. Narasimha[1], **Sanjeev C. Lingareddy**[2]
[1]Department of Computer Science and Engineering, Visvesvaraya Technological University, Bangalore, India
[2]Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Bangalore, India

Article Info

Article history:

Received Feb 6, 2023
Revised Sep 14, 2023
Accepted Nov 7, 2023

Keywords:

Cost
Hadoop
Hadoop map reduce
MapReduce
Video saliency

ABSTRACT

Video saliency detection is a rapidly growing subject that has seen very few contributions. The most common technique used nowadays is to perform frame-by-frame saliency detection. The modified spatio-temporal fusion method presented in this paper offers a novel approach to saliency detection and mapping. It uses frame-wise overall motion color saliency as well as pixel-based consistent spatio-temporal diffusion for its temporal uniformity. Additionally, a variety of techniques is advocated as a way to increase the saliency maps' overall accuracy and precision. The video is divided into groups of frames, and each frame temporarily goes through diffusion and integration in order to compute the color saliency mapping, as covered in the proposed method section. Then, with the aid of a permutation matrix, the inter-group frame is used to format the pixel-based saliency fusion, after which the features, or the fusion of pixel saliency and color information, direct the diffusion of the spatiotemporal saliency. The result is tested using five publicly accessible global saliency evaluation metrics, and it is determined that the proposed algorithm outperforms numerous saliency detection techniques with an improvement in accuracy margin. The robustness, dependability, adaptability, and precision are all demonstrated by the results.

Corresponding Author:

Sharada P. Narasimha
Department of Computer Science and Engineering, Visvesvaraya Technological University
Bangalore, India
Email: sharada_p2k22@redffimail.com

1. INTRODUCTION

The ability of the human eye to see is a remarkable feat of nature. The brain and eyes work together to form a potent system that is able to see 50 things per second and 10 million different varieties. The human eye can zoom in on certain elements of an image or video that are significant to us. By doing this, the brain filters out the irrelevant information and only retains that which is significant. A video is made up of a number of pictures, therefore the quantity of information that must be processed grows as does the impression of space.

In the technological realm, attempts have been made to copy or recreate this approach of image and video processing in a new manner. Itti's *et al.* model [1], which is thought to be the most popular model for stationary pictures, is one of the saliency models for stationary images that are currently accessible. Other theories include [2], which uses fourier transformation in a manner similar to phase spectroscopy, and [3], which employs frequency tuning for saliency detection. The use of the bottom-up visual attention process is what unites the aforementioned models. For instance, Achanta *et al.* [3] model generates the saliency map using a range of frequencies from the visual spectrum that emphasizes the key elements. The saliency map is then generated using the difference between Gaussians and a combination of the outputs from several bandpass filters. Then, using all of the low-level picture characteristics, feature maps are created [4], [5]. These maps are then added to the final saliency map output using visual nervous system-inspired concepts like winner take or inhibition of return. These are all still-image-focused and not video-focused. When watching videos, the texture element may not be as noticeable as it is in still images. Other saliency models and approaches are so required for videos; videos are collections of moving pictures, or frames that move in a certain order. In order to create a smooth motion and prevent the brain from being able to distinguish between each picture, there is a predefined frame rate. Videos may also be used to determine where one thing is in relation to another [6]. Video saliency is evidently far more complicated than picture saliency. Numerous studies have been conducted in this area, mostly using two techniques: one is the computing of a space-time saliency map, and the other is the computation of a motion saliency map [7]–[10]. By combining the concepts of static and dynamic saliency mapping to produce a space-time saliency detection model, which produced spatial-metrically mapped video saliency [11]. In order to get the motion patterns even for dynamic sceneries, Schütz *et al.* [12] presented a dynamic texture model.

In general, the majority of video saliency models start with bottom-up imagery since it can handle non-stationary movies. Additionally, motion information is seen as an additional saliency hint to aid in the detection of video saliency, and to achieve this, several radical saliency approaches combine the saliencies of motion and color. The fusion concept has been used by [13]–[15], but the outcome is low-level saliency. Nearly all of the most recent models maintain a brief smoothness in the saliency map of the results, which helps to increase accuracy. Even though [16], [17] employed global temporal cues to generate a low-level resilient saliency, these approaches suffer from error accumulation because they make use of the minimization of energy framework, which can manage saliency consistency across a temporal scale but results in false detections. As a topic with a little study, video saliency offers a lot of space for advancement, customization of models, and addition of restricted accuracy fall while ensuring temporal saliency consistency.

Modern picture saliency detection techniques are often used in video saliency algorithms as the fundamental saliency hints, however, in this research, the strategy is to employ simple low-contrast saliency without any high-level priors or restrictions. Additionally, incorporating the spatial-temporal gradient map prevents the hollow effect. The motion saliency and color saliency fusion are guided by the temporal-level global cue, which is used as the basis for appearance modelling. In order to assign high saliency values surrounding foreground objects and avoid taking into account the hollow effects, a spatial-temporal gradient definition is the recommended solution of the custom spatio-temporal fusion saliency detection approach. By making a number of changes to the saliency techniques, which aid in the fusion of motion and color saliencies, the efficiency and accuracy of the solution are increased. The first step in protecting the temporal smoothness is to create a temporal saliency correspondence using cross-frame super pixels. After that, the smoothness is used to further improve the saliency model's accuracy by applying a one-to-one spatial-temporal saliency diffusion.

2. RELATED WORK

This section will discuss the many research publications that served as inspiration for the bespoke spatio-temporal fusion saliency detection technique that is the suggested solution. As was already noted, visual saliency highlights the picture's most crucial elements. Due to the surge in traffic brought on by webinars, video streaming, and other factors, there has been exponential growth in video compression. Numerous video compression algorithms have been developed because of the desire for the highest possible video quality. These algorithms aim to squeeze more video into less memory while maintaining the same level of quality. Convolutional neural networks (CNN) have been used in this discipline as well. Learning-based video compression techniques have been surveyed [18], and the benefits and drawbacks of each technique have been examined. In order to advance the underdeveloped subject of video saliency, Borji [19] conducted research on the numerous deep saliency models, their standards, and datasets. The paper also discusses ways to address the disparities between algorithm-level and human-level saliency detection accuracy.

Three contributions are made in the meanwhile in [20]. First, they developed a brand-new benchmark called dynamic human fixation 1K (DHF1K), which aids in identifying fixations required for dynamic scene-free viewing. Then there is the attentive CNN with long sort term memory (LSTM) network (ACLNet), which adds a supervised attention mechanism to the CNN-LSTM architecture to facilitate quick end-to-end saliency learning. This makes it easier for the CNN-LSTM to concentrate on learning end-to-end saliency techniques more quickly for improved temporal saliency representation over subsequent frames. Their substantial testing on the three datasets DHF1K, Hollywood-2, and University of Central Florida (UCF) sports is the third contribution. The outcomes of the tests carried out were of the utmost and greatest significance for the advancement of the mentioned sector.

Achanta *et al.* [21] has provided a method to lessen the mistake produced in smooth pursuits (SPs), a significant sort of eye movement that is only seen while viewing dynamic situations. The approach makes use of manually annotated SPs, algorithmic fixation sites, fixation of salient SP locations, and saliency prediction using slicing CNN training. Then, using the methodologies currently in use, this solution model is evaluated on three datasets. Greater efficiency and precision are the ultimate results. Another model has been proposed [22] that predicts dynamic scene saliency using 3D convolutional encoder-decoder subnetworks. The decoder then expands the features in the spatial dimensions while simultaneously aggregating temporal data.

The outcome initially began with the extraction of spatial and temporal characteristics utilizing two subnetworks. The modern standard for video compression methods is the high-definition video compression system. The high efficiency video coding (HEVC) algorithms have been enhanced by [23] with the suggestion of a spatial saliency algorithm, that makes use of the idea of a motion vector. Based on a CNN, the motion estimate for each block is integrated during HEVC compression, and adaptive dynamic fusion occurs. Along with another technique to aid with rate distortion optimization, there is also an algorithm for a more flexible quantization parameter (QP) selection. The conditional random field (CRF) and saliency measure are combined in a novel salient object segmentation algorithm [24].

The resulting salient map is employed in CRF models employing segmentation approaches to build an energy minimization and recover clearly defined salient objects. It is created using a statistical framework and local feature contrast in color, lighting, and motion information. To determine visual saliency, Zhou *et al.* [25] also combines geographical and temporal data with statistical uncertainty measures. To create a single map, the two spatial and temporal maps are combined using a spatiotemporally adaptive entropy-based uncertainty weighting method. a contrast-based saliency in a pre-determined spatial and temporal environment is introduced in [26]. Ji *et al.* [27] discusses co-saliency detection utilizing cluster methods. Cluster saliency is calculated by combining the saliency maps from the single and many images using spatial, corresponding, and contrast metrics.

Another study uses a calculation of a robust geodesic measurement to produce saliency mapping [28], [29]. Using a super pixel-based approach, [30], [31] has contributed to the development of our suggested unique spatio-temporal fusion saliency detection technique. After being divided into superpixels, the picture is then subjected to adaptive color quantization. Based on the discrepancy between spatial distance and histograms, they then calculate inter-super pixel similarity. The super-pixel saliency map is created by first measuring the spatial and global contrast sparsity's, then integrating the results with inter-super pixels. The selection of the different saliency testing assessment metrics and procedures was aided here. It included references to the key studies and provided excellent metrics explanations that are easy to understand. The idea put forth by [32] is to combine temporal and spatial saliency using uncertainty weights based on entropy. In order to direct the fusion process, [16] turn to the mutual consistency between spatial and temporal saliency.

Even though fusion-based methods may determine the most reliable saliency cue from either the spatial or temporal saliency clues, failure scenarios nonetheless happen frequently when either the spatial saliency or the temporal saliency is off. The low-level saliency clues are often computed using spatial-temporal contrast-based approaches as opposed to fusion-based methods. [33], [34] to compute contrast-based saliency in a pre-defined spatial-temporal environment.

3. PROPOSED METHODOLOGY

The proposed approach is based on an improvised spatiotemporal saliency approach. The existing techniques are compared based on the generation of saliency maps in a frame-wise fashion. A contrast-based saliency technique is used here which is later modified and then the fusion of spatial and temporal saliencies takes place.

3.1. Initialization

A video saliency approach is applied here, which performs the identification in a compressed manner to reduce the number of computations associated with it. In Figure 1, a block diagram of the proposed approach is shown and divided into three segments. In the first phase, the saliency identification takes place, in the second phase the saliency modification takes place, in the third phase the fusion of spatio-temporal saliency takes for which the accuracy is enhanced by the smoothening method.

To enhance the saliency detection a part of the initial information is compressed in a frame-wise fashion. The streams and residual blocks are compressed via the vectors. The features are extracted from the data that is decoded. The reframed frames are then processed by which the spatial features are extracted the temporal features are extracted through the vectors. The extracted spatial features are compact, centred, and texture-based contrasted. In our proposed approach, these features are extracted and reframed in a compressed manner.

In the first phase the spatio-temporal saliency maps obtained are responsible to compare the reliability of the spatial map along with the temporal map, the relevant features are added to make it more robust, then the spatio-temporal saliency is identified. In the next phase, the saliency is modified henceforth making it robust to generate a saliency map for the long-term batch frame discussed in the later section. The issue of inconsistent saliency is solved by using the mechanism of fusion of spatial and temporal saliencies. Figure 1 shows the proposed architecture. The accuracy is boosted by the smoothening method for enhancing the accuracy by fusion of spatial and temporal saliencies. The low-level coherence problems are solved by using the model's alignment for non-stationary values and changing the background as a result to account for the camera's motion by introducing non-rigid variations.

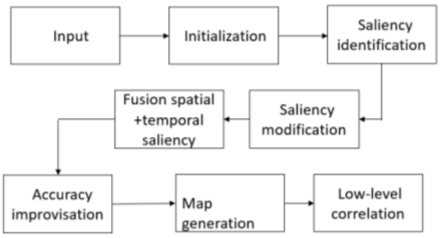


Figure 1. Proposed architecture

3.2. Video saliency identification

In a long-term frame-wise video sequence the video saliency is identified to find the salient object in each frame. The initial salient object is found by contrast-based saliency evaluation; a low-level saliency is adapted here based on contrast-based saliency computation. Here the long-term sequences are compressed in frame-wise short-term batches as $F_a = \{B_1, B_2, \dots\dots.., B_n\}$. Here B_x shows the $x - th$ video frame. In each video B_j the updated video batch the smoothening method is used that discards any irrelevant information and clustering is done to minimize the overhead to a great extent. The gradient map built is much stronger than the motion saliency, which combines the motion gradient with the colour gradient to obtain a Spatio-temporal gradient for low-level contrast evaluation showed by (1) by integrating colour and motion contrast.

$$CM = \left|\left|au, av\right|\right|_2 \odot \left|\left|\nabla(B)\right|\right|_2 \tag{1}$$

Here $\odot$ depicts the Hadamard product whereas $\nabla(B)$ shows the color gradient, $au\ and\ av$ dept the horizontal and vertical gradient. The motion contrast is computed by (2):

$$C_a = \sum_{H_b \in \mu_a} \frac{||A_a, A_b||_2}{||H_a, H_b||_2},\ \mu_a = \{\Upsilon + 1 \geq \left|\left|H_a, H_b\right|\right|_2 \geq \Upsilon + lc\} \tag{2}$$

In this equation H_a denotes the center-position of the super-pixel, whereas A denotes the two-way flow-gradient, μ_a shows the contrast for computation evaluated by the Euclidean distance in between the $a - th$ superpixel with Spatio-temporal gradient map denoted as $spatio - temporal$ as given in (3).

$$\Upsilon = \frac{lc}{||\Lambda(CM)||_0} \sum_{\Upsilon \in ||\Upsilon, a|| \leq lc} \left|\left|\Lambda\left(CM_{\ \Upsilon}\right)\right|\right|_0 \tag{3}$$

The lc value is set to initial local contrast computation in the range $lc = 0.5 \min\{width, height\}$, $\Lambda \rightarrow down\ sampling$, this function enhances the computational burden, and this method adopts the motion relevant data that controls the contrast computational range in different aspects. In the initial phase, the salient region is compared to that of the outer non-salient framework in avoiding the hollow effect. The contrast computational range is heavily focused on assuming the background regions that make the colour contrast computations a fail. In the second phase, the flow-based motion is assigned a larger-saliency value to the foreground object to discard the distance penalty. The colour saliency $color - sal$ is computed here by replacing it with the flow gradient along the red, green and blue (RGB) color value in (4). Here the RGB_a denote the RGB color value of the a − th super-pixel.

$$color - sal\ = \sum_{H_b \in \mu_a} \frac{||RGB_a),(RGB_b)||_2}{||H_a, H_b||_2} \tag{4}$$

$$color - sal_{\ x,a} \leftarrow \frac{\sum_{\Upsilon=x-1}^{x+1} \sum_{s_{\Upsilon,b \in \text{þ}\delta}} \exp\left(-||u_{x,a}, u_{\ \Upsilon,b}||1/\text{þ}\right) \cdot color - sal_{\ \Upsilon,b}}{\sum_{\Upsilon=x-1}^{x+1} \sum_{s_{\Upsilon,b \in \text{þ}\delta}} \exp\left(-||u_{x,a}, u_{\ \Upsilon,b}||1/\text{þ}\right)} \tag{5}$$

Henceforth, the $u_{x,a}$ is the average of the $x-th$ superpixel RGB color value for the $x-th$ frame, in the (6) $\left\|s_{x,a}, s_{Y,b}\right\|_2 \leq \partial$ is satisfied by this equation using ꞵ.

$$\partial=\frac{1}{p\times q}\sum_{x=1}^{q}\sum_{a=1}^{p}||\frac{1}{p}\sum_{a=1}^{p}W(CM_{x,a}),W(CM_{x,a})||_1 \tag{6}$$

$p,q = frame\,numbers$ in the present batch and the super-pixel in the present frame, $W(.)$ depicts the indicator function to choose the super-pixels with large CM values. ϵ Shows the filter strength whose value is set to 10. Here S_a shows the mean center coordinate for $x-th$ super pixel.

$$W\left(CM_a\right)=\begin{cases}S_a, & CM_a \leq \epsilon \times \frac{1}{p}\sum_{a=1}^{p} CM_a\\ 0, & else\end{cases} \tag{7}$$

Here in each batch frame-level the $a-th$ frame is smoothened and dynamically updated with the (7). The small range is much more effective than the tiny salient object, however in contrast it generates better effects that assigns a large smooth range for the purpose of a large salient object. The benefits here are introduced from ∂ which is fixed as mentioned in (8).

$$(1-\beta)\partial_{g-1}+\beta\partial_g \to \partial_g; \tag{8}$$

The color and motion saliency is fused resulting in pixel based saliency map. Here β depicts the learning weight whose value is set at 0.2. The $color-sal$ and $motion-sal$ value are fused to obtain the saliency value TT. Where $\odot$ denotes the Hadamard product, the fused saliency value is more efficient than the single-handed $color-sal$ or $motion-sal$ value as given in (9).

$$TT = color-sal \odot motion-sal\ ; \tag{9}$$

The saliency–maps enhance the accuracy of the rate depreciated. According to the performance evaluation, the enhancement is improved. The variation in the result leads to the introduction of a low-rank coherency-based saliency model for transmission.

3.3. Saliency modification

The short-term contrast information is not necessary to generate a robust saliency map, that integrates the long-term inter batch which is not sufficient to generate a robust saliency map; this suppresses the saliency degree for non-salient backgrounds. To appraise the salient background and non-salient background model, henceforth to accomplish the saliency detection of each frame batch. We shall use $BG_M \in \mathbb{R}^{3\times bg}$ and $FG_M \in \mathbb{R}^{3\times fg}$ to represent the background model and foreground appearance model, with $fg\ and\ bg$ being the sizes of their respective backgrounds, while their main aim is responsible for the $a-th$ super pixel's RGB history in all regions.

$$intra-cluster_a = \exp(\text{Ƥ} - |\alpha(motion-sal)-\alpha(color-sal)|);\ \lambda = 0.5 \tag{10}$$

$$inter-cluster_a = \alpha(\frac{\min||(RGB)_a,BG_M||_2\cdot\frac{1}{bg}\Sigma||RGB)_a,BG_M||_2}{\min||RGB)_a,FG_M||_2\cdot\frac{1}{fg}\Sigma||RGB)_a,FG_M||_2}) \tag{11}$$

When α is the upper bound value for the discrepancy degree. The inverse penalty is issued in between the motion and color saliency. The saliency here is modified based on two aspects. In the first phase both the salient object's focus and the non-salient background tend to remain unchanged in a restricted duration of consecutive frames, this facilitates modification of the saliency degree considering the previously developed foreground and background models. In the second phase, the colour saliency $color-sal$ is considered as the corresponding part of motion saliency whose main aim is to sharpen the minute detailed information of the salient object where the motion saliency shows the highest degree of saliency. According to the first step, the video-saliency is based on the extraction method performed in a batch-wise fashion, by considering the results previously developed from saliency identification. The background and foreground model is fixed by incorporating the previously identified results. This necessarily utilizes the RGB colour value to adjust the colour saliency value.

3.4. Fusion of spatial and temporal saliency on low-level

The saliency obtained is much better however, there exist many issues with the saliency value is not consistent. In this section, the accuracy obtained via the saliency map is boosted while enhancing the smoothening based on the fusion of spatial and temporal saliency. The problem on the low-level which aims at decomposing the input matrix V into low-rank part O and sparse part as G as $V = O + G$. The problem is formulated as (12) and (13):

$$\min_{O,G} Ѓ\left|\left|O\right|\right| + \left|\left|G\right|\right|_o \qquad subj = V = O + G \tag{12}$$

$$\min_{O,G} Ѓ\left|\left|O\right|\right|_* + \left|\left|G\right|\right|_1 \qquad subj = V = O + G \tag{13}$$

In (12) shows a non-convex NP-hard problem solved by relaxing the convex envelope as shown in (13). ||·||∗ shows the nuclear representation solved via the principal component analysis. The optimal solution is division into two segments. In (14) single peak value for low-level estimation.

$$G \leftarrow sign(V - O - G)\left[|V - O - G| - Ѓß\right]_+ \tag{14}$$

$$O \leftarrow J\left[\Sigma - Ѓß\right]_+ K, (K, \Sigma, J) \leftarrow singvaldec(L) \tag{15}$$

Here (L) depicts the Lagrange multiplier whereas Ѓ and ß denotes the low-level value and the sparse parameter. The approach formulated here uses only the sparse component whereas thereby solving the problems associated with it. However, as these low-rank problems adopt the alignment for non-stationary values for the model and hence the transformation of the background to handle the motion of the camera by introducing non-rigid variations. This easily results in many problems; this method adopts various low-level constraints to minimize the non-salient backgrounds to enhance the salient foregrounds, which is more accommodating to handling the restrictions of non-salient videos. This method accommodates the low-rank revealing and background modelling. Because the overlapped foreground region is slow and easily modified as a low-level background model.

Parallel, to minimize the problems due to inappropriate optical flow the super-pixels enclosed in the given region's where the foreground region is located and feature subspace of a frame k is traversed as $vZ_j = \{TT_{g_{j,1}}, LL_{g_{j,2}}, \ldots.. LL_{g_{j,n}}\}$ and thus for the entire frame group we get $vB_Y = \{vI_1, vI_2, \ldots., vI_n\}$. This way the foreground model is evaluated as (16):

$$E_{H_a} = [\textstyle\sum_{j=1}^{q} TT_{g_{j,a}} - \frac{\alpha}{q \times p} \sum_{j=1}^{q} \sum_{a=1}^{p} TT_{G_{j,a}}]_+ \tag{16}$$

In (15) α denotes the reliability measure for two feature sub-spaces in eq 15 traversed by TT_g and RGB color value for $VS = \{gs_1, gs_2, \ldots., gs_n\} \in J^{3p \times n}$ $gs_a = \{vec\left(R_{a,1}, G_{a,1}, B_{a,1}, \ldots., R_{a,p}, G_{a,p}, B_{a,p}\right)\}^J$ and $D_E = vec\left(TT_{g_1}\right), \ldots. vec(TT_{g_n}) \in J^{v \times n}$. This leads to a one-to-one relevance and then pixel-based saliency mapping fusion that is consumed by an entire group of frames. VS upon D_E leads to the foreground salient model and considering this the problem is solved by the issue of an alternate approach.

$$\min_{V_{mk}, G_{ca}, \tau, H \odot \tau} \left|\left|V_c\right|\right|_* + \left|\left|O_x\right|\right|_* + \left|\left|H + \tau\right|\right|_2 + \mu_1\left|\left|G_c\right|\right|_1 + \mu_2\left|\left|G_a\right|\right| \tag{17}$$

$$s.t\ V_c = O_c + G_c,\ V_s = O_s + G_x,\ V_c = VS \odot, \tau, V_x = DE \odot \tau, \tag{18}$$

$$\tau = \{U_1, U_2, \ldots., U_n\}, U_a \in \{0,1\}^{p \times p}, U_a 1^J = 1 \tag{19}$$

Here the pixel-embedded features over the color and saliency feature subspaces are represented by O_c, O_x variables, τ is the permutation matrix. this assists in accommodating the super-pixel correspondence.

4. RESULT ANALYSIS

In this section the results are evaluated by comparing our proposed model with the existing techniques such as operational block description length (OBDL) algorithm, dynamic adaptive whitening saliency (AWS-D) algorithm, object-to-motion convolutional neural network two layer long short-term memory

(OMCNN-2CLSTM) algorithm, attentive convolutional (ACL) algorithm, saliency aware video compression (SAVC) algorithm, and XU. 10 video sequences were taken in each of the three distinct resolutions of 1,920×1,080, 1,280×720, and 832×480 for the final comparison and evaluation, Table 1 demonstrates this. The findings are then displayed using five evaluation metrics: area under ROC curve (AUC), similarity or histogram intersection (SIM), pearson's correlation coefficient (CC), normalized scanpath saliency (NSS), and kullback-leibler divergence (KL). The database used here is a high-definition eye-tracking database by using an appropriate open-source GitHub repository [35]. These algorithms widely focussed via a great common intermediate format (CIF) resolution also focussed on HD compatible video. Table 2 in Appendix shows the results of all 5-algorithm used for video saliency.

Table 1. Comparison within three distinct resolutions

Type	Resolution	Name	Frame Rate (Hz)
A	1920×1080	BasketballDrive	50
		Kimono 1	24
		ParkScene	24
		Johnny	60
B	1280×720	KristenAndSara	60
		FourPeople	60
		vidyo3	60
		vidyo4	60
C	832×480	BasketballDrill	50
		RaceHorses	30

4.1. Result section

Ln this section the results are evaluated by the above-mentioned methods OBDL, AWS-D, OMCNN-2CLSTM, ACL, SAVC, Xu algorithm existing system, and our proposed approach. Similar to the existing system, five widely used saliency evaluation metrics have been employed: AUC, SIM, CC, NSS, and KL.

4.1.1. Area under ROC curve

AUC, is a saliency model evaluation metric where OBDL has an AUC value of 0.6413, AWS-D has an AUC value of 0.6635, OMCNN-2CLSTM has a value of 0.7322, ACL has 0.7673, SAVC method has least AUC value of 0.5844 similar value is achieved by XU of 0.5881 the existing system has an AUC value of 0.7334 whereas the proposed system performs better and has an AUC value of 0.7354. Table 3 and Figure 2 displays the AUC value comparison.

Table 3. AUC value comparison

Method	AUC
OBDL	0.6413
AWS-D	0.6635
OMCNN-2CLSTM	0.7322
ACL	0.7673
SAVC	0.5844
XU	0.5881
Existing system	0.7334
Proposed system	0.7354

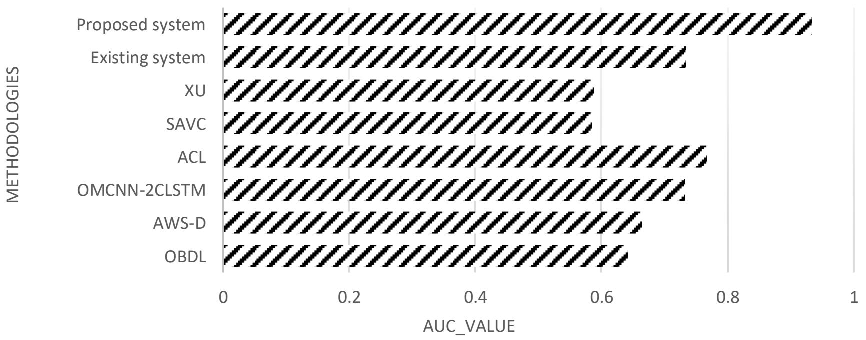


Figure 2. AUC value comparison

4.1.2. Similarity or histogram intersection

SIM, for this metric the OBDL methodology has a value of 0.2982 and AWS-D has a value of 0.3154, OMCNN-2CLSTM has a SIM value of 0.3664, ACL has a value of 0.3614, SAVC has a value of 0.2688 and XU has 0.305 value, the existing system has a value of 0.3751. The proposed system performs better than the existing system and achieves a value of 0.407378. Table 4 and Figure 3 displays the SIM value comparison.

Table 4. SIM value comparison

Method	SIM
OBDL	0.2982
AWS-D	0.3154
OMCNN-2CLSTM	0.3664
ACL	0.3614
SAVC	0.2688
XU	0.305
Existing system	0.3751
Proposed system	0.407378

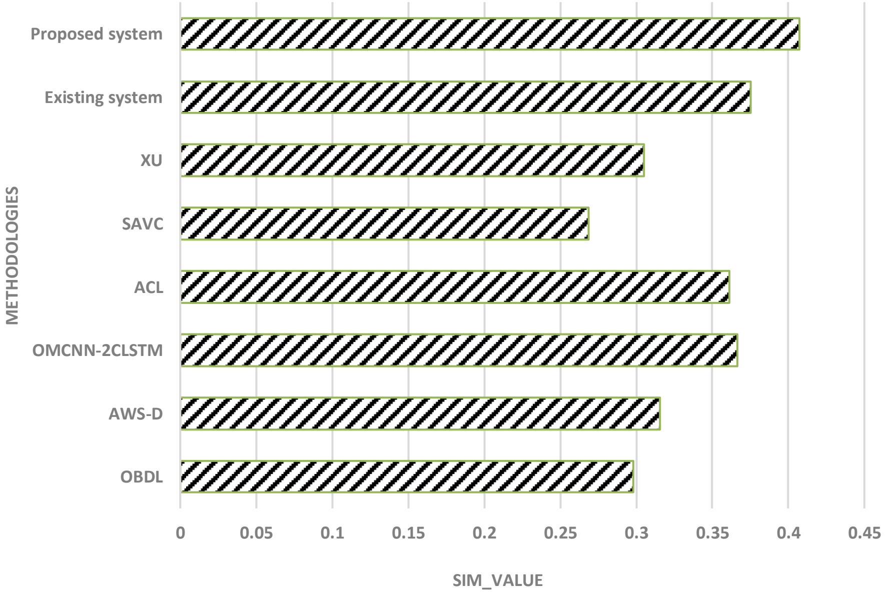


Figure 3. SIM value comparison

4.1.3. Pearson's correlation coefficient

CC, in Table 3 the OBDL methodology has a value of 0.2253, AWS-D methodology has a value of 0.2663, OMCNN-2CLSTM methodology gives a CC value of 0.4501 and ACL methodology achieves a value of 0.3774, SAVC methodology gives a value of 0.1248, XU method gives 0.2663 CC value the existing system achieves an value of 0.387. The proposed system gives better results in comparison with the existing system and attains a value of 0.44391. Table 5 and Figure 4 shows the CC value comparison.

Table 5. CC value comparison

Method	CC
OBDL	0.2253
AWS-D	0.2663
OMCNN-2CLSTM	0.4501
ACL	0.3774
SAVC	0.1248
XU	0.2663
Existing system	0.387
Proposed system	0.44391

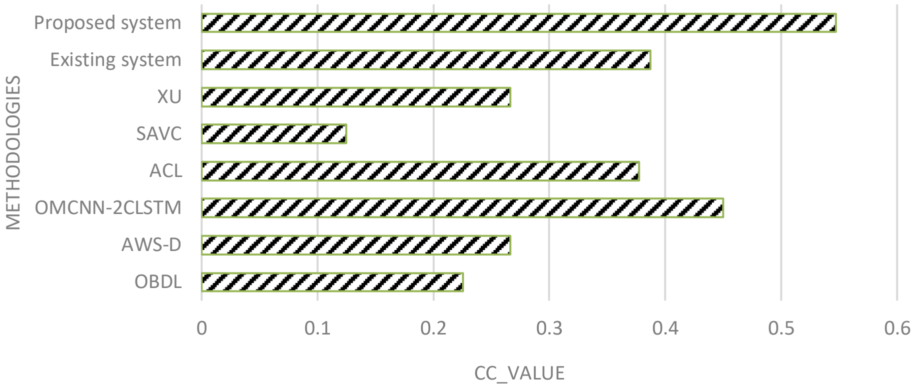


Figure 4. CC value comparison

4.1.4. Normalized scanpath saliency

NSS in Table 4 the OBDL methodology has a value of 0.297, AWS-D methodology has a value of 0.4768, OMCNN-2CLSTM methodology gives an NSS value of 0.5585 and ACL methodology achieves a value of 0.5005, SAVC methodology gives a value of 0.1889, XU method gives 0.2854 NSS value the existing system achieves a value of 0.5674. The proposed system gives better results in comparison with the existing system and attains a value of 1.000138. Table 6 and Figure 5 shows the NSS value comparison.

Table 6. NSS value comparison

Method	NSS
OBDL	0.297
AWS-D	0.4768
OMCNN-2CLSTM	0.5585
ACL	0.5005
SAVC	0.1889
XU	0.2854
Existing system	0.5674
Proposed system	1.000138

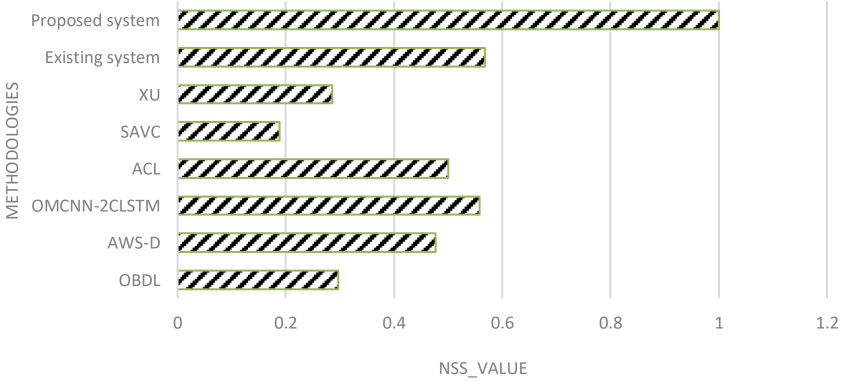


Figure 5. NSS value comparison

4.1.5. Kullback-leibler divergence

KL in Table 5 the OBDL methodology has a value of 3.4642, AWS-D methodology has a value of 2.0191, OMCNN-2CLSTM methodology gives a KL value of 2.82 and ACL methodology achieves a value of 3.0642, SAVC methodology gives a value of 2.0191, XU method gives 1.5098 NSS value the existing system achieves a value of 2.4921. The proposed system gives less result only for KL metric in comparison with the existing system and attains a value of 0.862871. Table 7 and Figure 6 displays the KL value comparison.

Table 7. KL value comparison

Method	KL
OBDL	3.4642
AWS-D	1.7144
OMCNN-2CLSTM	2.82
ACL	3.0642
SAVC	2.0191
XU	1.5098
Existing system	2.4921
Proposed system	0.862871

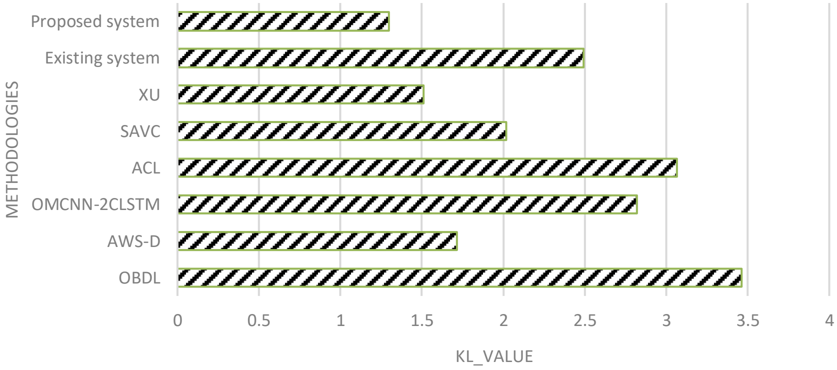


Figure 6. KL value comparison

4.2. Comparative analysis

The saliency video compression mechanism is evaluated using the metrics AUC, SIM, CC, NSS, and KL. A comparative analysis is carried out by comparing our proposed system with the existing system and the percentage improvisation for each metric is shown in the Table 8. For AUC metric the existing system attains a value of 0.7334 whereas the proposed system has a value of 0.7354 and the percentage improvisation from existing system to the proposed system is 0.2723%. For SIM metric the existing system attains a value of 0.3751 whereas the proposed system has a value of 0.407378 and the percentage improvisation from existing system to the proposed system is 8.23%. For CC metric the existing system attains a value of 0.387 whereas the proposed system has a value of 0.44391 and the percentage improvisation from the existing system to the proposed system is 13.69%. For the NSS metric the existing system attains a value of 0.5674 whereas the proposed system has a value of 1.000138 and the percentage improvisation from the existing system to the proposed system is 55.2%. However, our proposed approach does not perform well with the KL metric henceforth the comparative analysis is not done for it, however, our proposed model outperforms the existing system for various other metrics.

Table 8. comparative analysis for various metrics

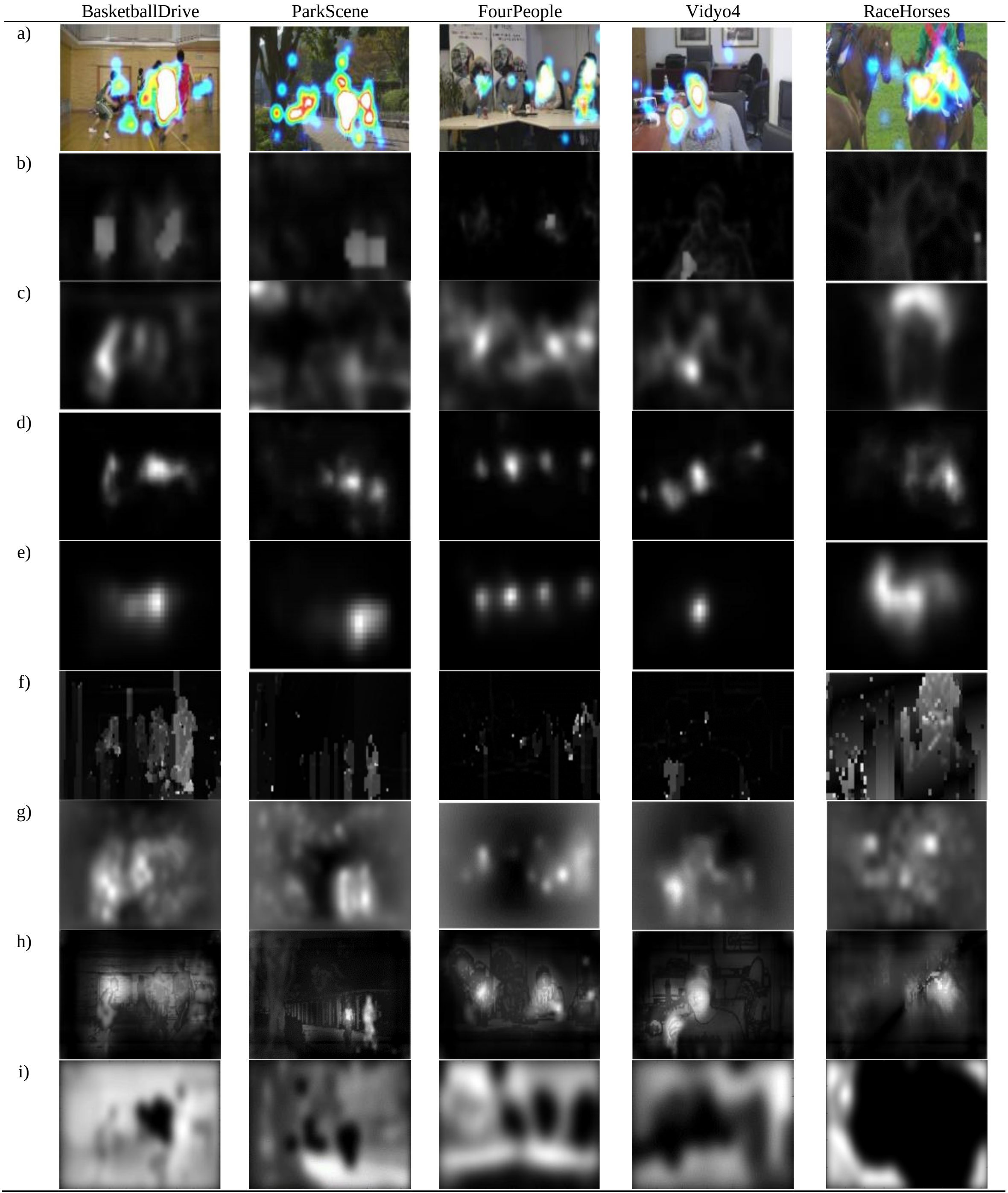
Improvisation	AUC	SIM	CC	NSS
Existing system	0.7334	0.3751	0.387	0.5674
Proposed system	0.7354	0.407378	0.44391	1.000138
Improvisation in percentage	0.2723 %	8.23%	13.69%	55.2%

5. CONCLUSION

In contrast to the most recent state-of-the-art saliency identification techniques, this research introduces a modified spatio-temporal fusion video saliency detection method that is more accurate and precise. Simple calculations have been modified in several ways to address the issues with colour contrast computation, including the fusion aspect of saliency, which has been improved to boost both motion and colour values, and spatio-temporal of pixel-based coherency, which has been improved for temporal scope saliency exploration. The final solution had been evaluated against a sizable database in order to understand its reliability and effectiveness. The final result has also been contrasted with other radical saliency mapping techniques, and it has been found that the proposed approach has higher accuracy and precision. With all these changes, the performance of our proposed modified spatio-temporal fusion video saliency detection approach has significantly improved, giving the field of video saliency a new cause for optimism. Given the rareness of existing research, this technique will be useful for people who wish to carry out additional saliency detection studies.

APPENDIX

Table 2. The following results for saliency algorithms used; a) fixation maps, b) OBDL, c) AWS-D, d) OMCNN-2CLSTM, e) ACL, f) SAVC, g) XU, h) base paper, and i) proposed algorithm



REFERENCE

[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 *2008 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 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*, IEEE, Jun. 2009, 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] S. H. Sreedhara, V. Kumar, and S. Salma, “Efficient big data clustering using adhoc fuzzy C means and auto-encoder CNN,” in *Lecture Notes in Networks and Systems*, vol. 563, 2023, pp. 353–368, doi: 10.1007/978-981-19-7402-1_25.

[6] L.-J. Li and L. F.-Fei, “What, where and who? Classifying events by scene and object recognition,” in *2007 IEEE 11th International Conference on Computer Vision*, IEEE, 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. G.-Dugué, “Spatio-temporal saliency model to predict eye movements in video free viewing,” in *European Signal Processing Conference*, 2008, pp. 1-5.

[9] Y.-F. Ma and H.-J. Zhang, “A model of motion attention for video skimming,” in *Proceedings International Conference on Image Processing*, IEEE, 2002, pp. 129–132, 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*, IEEE, 2007, pp. 1073–1076, 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*, IEEE, 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*, IEEE, Jun. 2014, pp. 3358–3365, doi: 10.1109/CVPR.2014.429.

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

[15] Y. Fang, Z. Wang, and W. Lin, “Video saliency incorporating spatiotemporal cues and uncertainty weighting,” *Proceedings - IEEE International Conference on Multimedia and Expo*, 2013, doi: 10.1109/ICME.2013.6607572.

[16] W. Wang, J. Shen, and F. Porikli, “Saliency-aware geodesic video object segmentation,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3395–3402, 2015, doi: 10.1109/CVPR.2015.7298961.

[17] L. Huang, P. Yan, G. Li, Q. Wang, and L. Lin, “Attention embedded spatio-temporal network for video salient object detection,” *IEEE Access*, vol. 7, pp. 166203–166213, 2019, doi: 10.1109/ACCESS.2019.2953046.

[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, 2021, doi: 10.1109/TPAMI.2019.2924417.

[21] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua and S. Süsstrunk, “SLIC superpixels,” *EPFL Technical Report*, pp. 1-15, 2010.

[22] D. Zhang, O. Javed, and M. Shah, “Video object segmentation through spatially accurate and temporally dense extraction of primary object regions,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2013, pp. 628–635, doi: 10.1109/CVPR.2013.87.

[23] 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.

[24] 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.

[25] 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.

[26] Z. Zeng, T. H. Chan, K. Jia, and D. Xu, “Finding correspondence from multiple images via sparse and low-rank decomposition,” in *Computer Vision – ECCV 2012*, Springer, 2012, pp. 325–339, doi: 10.1007/978-3-642-33715-4_24.

[27] P. Ji, H. Li, M. Salzmann, and Y. Dai, “Robust motion segmentation with unknown correspondences,” in *Computer Vision – ECCV 2014*, vol. 8694, Springer, 2014, pp. 204–219, doi: 10.1007/978-3-319-10599-4_14.

[28] 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)*, IEEE, 2005, pp. 1016–1021, doi: 10.1109/CVPR.2005.264.

[29] 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.

[30] 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, 1957, doi: 10.1137/0105003.

[31] Z. Liu, L. Meur, and S. Luo, “Superpixel-based saliency detection,” in *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, IEEE, Jul. 2013, pp. 1–4, doi: 10.1109/WIAMIS.2013.6616119.

[32] 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.

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

[34] V. Leboran, A. G. -Diaz, X. R. F. -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.

[35] F. Zhang and Spzhubuaa, “VED100-a video based eye tracking dataset on visual saliency detection,” *GitHub,* 2018. [Online]. Available: https://github.com/spzhubuaa/VED100-A-Video-Based-Eye-Tracking-Dataset-on-Visual-Saliency-Detection

BIOGRAPHIES OF AUTHOR

Prof. Sharada P. Narasimha received the B.E. Degree from BIET, Belagavi, India, in 2003, M.Tech. Degree in CSE from SJCIT, India, in 2013. She is pursuing towards her Ph.D. in VTU-RC at SVCE. She is having total 10+ years of work experience in teaching/research field and 4 years in Industry. Editor of IJRP, IJLTEMAS, she is currently working as an Assistant Professor in Department of Computer Science and Engineering in Sri Venkateshwara College of Engineering, Bengaluru. She has organized and conducted FDP/SDP/webinars/conferences. Her areas of research are image processing, wireless networks, data communications, game theory, network security, and computer networks, IOT, MEMS, and embedded systems. She can be contacted at email: sharada_p2k22@redffimail.com.

Dr. Sanjeev C. Lingareddy received his Ph.D. in the year of 2012 from JNTU, Hyderabad and currently working as Professor and Head for the Department of Computer Science and Engineering at Sri Venkateshwara College of Engineering, Bengaluru. He has 24 years of rich experience in the academics and 7 years of research experience. He has published more than 25 research articles in International Journals. He is a Member of Indian Society for Technical Education (MISTE) and an active member in many technical events. His research area includes wireless sensor network, wireless security, cloud computing, and cognitive network. He can be contacted at email: sclingareddy@gmail.com.