Video saliency detection using modified high efficiency video coding and background modelling

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 Venkateshwara College of Engineering, Bangalore, India

Article Info

Article history:

Received Feb 8, 2023
Revised Nov 7, 2023
Accepted Dec 22, 2023

Keywords:

Background modelling
High efficiency video coding
Image saliency
Video input
Video saliency

ABSTRACT

Video saliency has a profound effect on our lives with its compression efficiency and precision. There have been several types of research done on image saliency but not on video saliency. This paper proposes a modified high efficiency video coding (HEVC) algorithm with background modelling and the implication of classification into coding blocks. This solution first employs the G-picture in the fourth frame as a long-term reference and then it is quantized based on the algorithm that segregates using the background features of the image. Then coding blocks are introduced to decrease the complexity of the HEVC code, reduce time consumption and overall speed up the process of saliency. The solution is experimented upon with the dynamic human fixation 1K (DHF1K) dataset and compared with several other state-of-the-art saliency methods to showcase the reliability and efficiency of the proposed solution.

Corresponding Author:

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

1. INTRODUCTION

The human eye is a complex organ, the way it works with the brain to filter and analyses the necessary components in the image it sees has perplexed scientists for ages, and many have tried to replicate it using algorithms and computations. Using the process done in the brain, researchers have tried to develop methods to pick out those areas of interest from the given image, just like the human visual system. With the inclusion of deep learning technology, there has been a significant rise in methods of image saliency detection with remarkable accuracy when tested on large-scale static gaze datasets such as the silicon dataset [1]. However, there have been several types of research done in the field of image saliency detection; it is quite challenging to produce the same effect of dynamic fixation prediction with moving images or videos. Video saliency has a great role in video compression, captioning, object segmentation and so on. This has led to the classification of saliency into two models, namely salient object detection and human eye fixation prediction. The input is also of two types, dynamic and static saliency models. Static models, as the name suggests, have images as their input and likewise, dynamic models take video input.

The inspiration for this paper has stemmed from various research papers based in a similar field. This paper have a significant impact on the world of saliency [2], [3]. These two papers have used the difference in the features between the surrounding and central patches to estimate visual saliency. This is another research that attempts to detect saliency by representing a combinational block, based on random walk models, of all neighboring blocks [4]–[6] has a unique technique that involves graphs and is named as graph-based visual saliency. It includes the formation of activation maps on certain features followed by normalization. It has an amazing receiver operating characteristic (ROC) value of about 98%. They have a similar approach to the problem with difference being that it uses random walk models on a graph to imitate the eye movements [7]. Their first step was to extract intensity, colour and compactness features, construct a fully connected graph and then the proposed algorithm computes the stationary distribution of the Markov chain on the graph as a saliency map. They have used another method of saliency detection using spectral features of an image. Exploits the features of images like luminescence and colour, which helps in reducing computational complexity and gives accurate results [8]–[10]. The researchers have used a base method involving regional application in saliency detection. In this, the input image is first segregated into different regions for saliency levels to be applied to each of them uses global contrast features with spatial weighted coherency, while [11]–[15] uses robust background measures along with a principled optimization framework to integrate all low-level maps to create a final clean and uniform saliency map. All the above algorithms are used for still images and these help in creating algorithms for video saliency detection. To modify them to be able to accurately detect visual saliency, we would need motion information to imitate the human eye's perception of movement. They have quaternion representation using the features of images like colour, intensity and motion and employment of phase spectrum of quaternion Fourier transform. This methodology involves discriminant center surround hypothesis. It combines colour orientations, and spatial and temporal saliency by taking summation of the absolute difference between temporal gradients of central and surrounding regions [16]. They made use of feature extraction from the partially decoded data. It uses global and local spatiotemporal (GLST) features [17]. The compressed video bitstream is partially decoded to obtain discrete cosine transform (DCT) coefficients and motion vectors and then GLST features are extracted. Then the spatial and temporal maps are generated and fused to get the result. This paper uses random walk with restart methodology. They figure out temporal saliency distribution using motion distinctiveness, abrupt change and temporal consistency [18]. Then it is used as restarting distribution and steady-state distribution is used to find spatiotemporal saliency.

All these researches and experiments tell us that many state-of-the-art methods are available in the uncompressed domain. Since videos and images are generally sent in a compressed format, these conventional algorithms do not perform well in these situations. The only way for them to work effectively on the available data is to fully decode the data but this increases time consumption and the complexity of the code. There has been some research to solve this problem [19]. Has tried to improve the DCT-domain transcoder or deterministic discrete-time (DDT), by proposing a fast extraction method for partial low-frequency coefficients in DCT domain motion compensation operation (DCT-MC). Zhang *et al.* [20] is redesigned to exploit the low-level compressed domain features from the bitstream. They uses object recognition for fast saliency detection. Colour clustering and region merging is based on spatiotemporal similarities, pixel edge extraction and regional classification [21]. They have similar video saliency detection methods [22]–[25]. They, have come across several methods for bettering the saliency area. One of them introduced the G-picture methodology, which meant that reference will be maintained, probably a second frame reference, for reducing the complexity of the high efficiency video coding (HEVC) algorithm, then there is the usage of a quantization parameter for quantizing the G-picture (ground) and with the employment of background reference prediction (BRP) and background difference prediction (BDP). Even small coding blocks called coding units were introduced to lower complexity and increase efficient compression [26]–[30]. However, all these works were in different times and different regions of work. We have tried to incorporate all these modifications to come up with a solution that not only reduces complexity but also helps in input size flexibility with reduced time consumption and better compression precision, accuracy and efficiency [31]–[33].

In this study, a modified version of the HEVC method is suggested that makes use of backdrop modeling with a hierarchical prediction structure (HPS). It consists of two parts. The first is the modification of the reference frame used, making G-picture the fourth reference frame rather than the second as stated in other research papers, quantizing it with a relatively smaller valued parameter, and adding coding blocks for less complexity. The division of each coding block into $\mathrm{F_G}, \mathrm{B_G}$ and $\mathrm{H_G}$ is the second element. Depending on the information included in the G-picture, each of these elements is sped up in a unique way. To avoid further coding and calculation, another alteration is included in which the coding block portioning is halted early.

There are a total of five sections in this essay. The introduction is covered in the first section, and the related works for this paper are listed in the second. The third section covers the mathematical and coding components of the suggested system, and the fourth section displays the outcomes of the tests done using the dataset dynamic human fixation 1K (DHF1K). The paper is then concluded in the fifth portion.

2. LITERATURE SURVEY

This section will provide a quick overview of the numerous studies and tests that have aided in the development of our solution. We now have a better iteration of the HEVC method, starting with [34]–[37], in which the perceptual redundancy has been decreased for higher compression value. With the use of a convolutional neural network, this suggested technique combines the motion estimation results from each block during the compression phase and employs adaptive dynamic fusion for the saliency map. The fundamental element of this suggested algorithm is the application of the spatiotemporal algorithm. The next one is a survey that has assisted in grouping and selecting the appropriate database as well as the modification approach for our suggested solution. They provides an up-to-date overview of all the video compression research along with its milestones [38]–[40]. It is done for conventional codec adaption along with learning-based end-to-end and their advantages and disadvantages. In their conclusion, the computation complexity is an issue that needs solving at the earliest available opportunity. This paper is another survey about the different saliency models available and what are the drawbacks that have led to insufficient accuracy and precision in compression [41]. The researcher has provided insight into the different ways the various saliency models can try to mimic the actual process of the human eye and brain.

This has helped in making the right modification to our algorithm with actual practical comparisons. The main dataset that has been used for the proposed solution's experiment as well as the dataset of the base reference that is used for comparison of our results [27], [42]. The dynamic human fixation 1K, often known as DHF1K, forecasts fixations when viewing dynamic scenes. With 1000 high-definition, diverse video clips taken by 17 observers while wearing eye trackers. Attentive convolutional neural network (ACLNet)-long sort-term memory (LSTM) network is a cutting-edge video saliency approach that has also been proposed. Additionally, it has contrasted its findings with those of other techniques using various datasets, including Hollywood-2 and University of Central Florida (UCF) sports. It was one of the quickest approaches up to this point. They have given us knowledge on hyper saliency [43]. Convolutional neural networks are trained using manual algorithmic annotations of smooth pursuits, and the findings are developed with the aid of 26 dynamic saliency models that are freely available online. Here another study that has aided in algorithm development? For prediction in dynamic scenarios, they have devised a brand-new 3-dimensional (3D) convolutional encoder-decoder architecture [41]. The encoder has two subnetworks that separate the spatial and temporal components of each frame and then fuse them. The decoder then aggregates temporal data and enlarges the features in spatial dimensions. It is tested on the DHF1K dataset after receiving end-to-end training. This is another survey of various video saliency methods available in today's world that employs deep learning and has tried its level best to reach the human level of eye tracking movements and feature detection [44].

They provides a no-reference bitstream human vision system (NRHVS) based video quality assessment (VQA) [45]. The saliency maps are generated by extracting the features from the HEVC bitstream and then a visual memory model is created using saliency map statistics. The support vector regression pipeline helps in learning the approximate video quality. VS-video saliency (DeepVS2.0) is a video saliency prediction approach based on deep neural networks [39]. It has aided in comparing our outcomes and evaluating how we did against other cutting-edge techniques. In order to create the intra-frame saliency map, it has presented an object-to-motion convolutional neural network (OM-CNN) that learns spatiotemporal properties. Then, using the OM-CNN extracted features, a convolutional LSTM network is created to enable inter-frame saliency. Our baseline reference is [46]. For different levels of the 3D convolutional backbone for the video saliency mapping, it uses its spatiotemporal self-assessment (STSANet) model [47]–[50]. In order to integrate many levels with context in semantic and spatiotemporal subspaces, attentional multi-scale fusion (AMSF) is used.

3. PROPOSED SYSTEM

3.1. Optimizing low-delay hierarchical prediction structure efficiently

In this part, we will briefly discuss the constituents of the low delay HPS of the HEVC test model. They are namely two components. One is called hierarchal quantization (HQ), which uses the data of the last frame and other prioritized frames from the last three short groups of frames, and the other is called hierarchal reference (HR). Where the quantization parameter of each important frame is the same as two less than its next image while the quantization parameter of the middle image in the short group of frames is equivalent to one more than the important frame's value. To optimize it, we need to replace the fourth reference frame with the G-Picture (generated using a general running less complex algorithm). This will remain as a long-term reference. For this, we shall use the Lagrange rate-distortion optimization and this helps in evaluating the rate distortion (RD) cost C. Where Q denotes the quality of reconstructed video about the original, η denotes the number of bits and μ denotes the Lagrange multiplier. There will be m input frames, and let $(\mathrm{I_i}, \mathrm{p})$ represent the rate-distortion cost of encoding the i-th picture $(\mathrm{I_i})$. p will represent the coding units' quantization parameter using a cost function.

$$\mathrm{C} = \mu\eta + \mathrm{B} \tag{1}$$

$$\mathrm{C} = \sum_{\mathrm{i}}^{\mathrm{m}} \Psi(\mathrm{I_i}, \mathrm{p}) = \sum_{\mathrm{i}}^{\mathrm{m}} \sum_{\mathrm{r}} \tau(\mathrm{p}, \mathrm{I_{i,r}}, \mathrm{Q_{i,r,p}}\ \mathrm{U_{i,r,p}}) \tag{2}$$

Where $U_{i,r,p}$ represents the motion vectors and $Q_{i,r,p}$ is the data prediction quantized with p. With a smaller p', it provides a better reference for a images $(I_{j+1} \sim I_{j+a})$. Assuming that there are n_{j+1} coding blocks for I_{j+1} for indexes $e(j+1,1) \sim e(j+1,n_{j+1})$ for better reference I_j but the other coding blocks q_{j+1} for indexes $t(j+1,1) \sim t(j+1,q_{j+1})$ cannot do so. This is similar for $I_{j+2} \sim I_{i+a}$ with n_{j+s} and this has better prediction than coding blocks indexed by $t(j+s,1) \sim t(j+s,q_{j+s})$ and $e(j+s,1) \sim e(j+s,n_{j+s})$. The new costing equation comes out to be as shown in (3).

$$C' = T_1 + T_2 + T_3 + T_4 + T_5 \tag{3}$$

As it can be deciphered from (4), T_1 give the rate-distortion cost before I_j is used, T_2 is the rate-distortion cost after using p' for encoding, T_3 is costing for coding blocks $I_{j+1} \sim I_{j+a}$, T_4 is the rate-distortion cost for all the combined rates of the coding blocks for modified I_j and T_5 is cost for I_{j+a}.

$$T_1 = \sum_{i=1}^{j-1} \Psi(I_i, p),$$

$$T_2 = \Psi(I_i, p'),$$

$$T_3 = \sum_{i=j+1}^{j+a} \sum_{l=1}^{Q_i} \tau(p, I_{i,t(i,l)}, Q_{i,t(i,l),p,} U_{i,t(i,l),p}),$$

$$T_4 = \sum_{i=j+1}^{j+a} \sum_{r=1}^{n_i} \tau(p, I_{i,e(i,l)}, Q_{i,e(i,l),p',} U_{i,e(i,l),p'}),$$

$$T_5 = \sum_{i=j+a+1}^{m} \Psi(I_i, p). \tag{4}$$

The modified costing equation using p instead are shown in (5). Now calculating the difference between (5) and (6), we get (7).

$$C = T_1 + X + T_3 + Y + T_5 \tag{5}$$

$$X = \Psi(I_j, p),$$

$$Y = \sum_{i=j+1}^{j+a} \sum_{r=1}^{n_i} \tau(p, I_{i,e(i,l)}, Q_{i,e(i,l),p,} U_{i,e(i,l),p}). \tag{6}$$

$$C - C' = (Y - T_4) - (T_2 - X) \tag{7}$$

In T_4, the term $Q_{i,e(i,l),p',}$ has lesser quantization loss than the term $Q_{i,e(i,l),p,}$ $(i = j+1 \sim j+a, l = 1 \sim n_i)$ in Y due to this, the inequality is satisfied is shown in (8).

$$Y - T_4 = \sum_{i=j+1}^{j+a} \sum_{r=1}^{n_i} (\tau(p, I_{i,e(i,l)}, Q_{i,e(i,l),p,} U_{i,t(i,l),p}) - \tau(p, I_{i,e(i,l)}, Q_{i,e(i,l),p',} U_{i,e(i,l),p'})) > 0 \tag{8}$$

Thus, the conclusion can be stated as - "for a large a in rate-distortion cost, C' that satisfies the equation $Y - T_4 > T_2 - X$, then $C - C' > 0$." There can be several conclusions drawn from the analysis of the equations above. If we frequently choose an image as a reference for the next batch of pictures, then the quantization parameters must be selected in such a way that the values are relatively smaller also on extending this conclusion. We can say that the G-picture, as it is taken as a long-term reference, must be quantized at a value lesser than the quantization parameter.

For the above conclusion to work well there must be the availability of a large number of pixels that have the same features as the G-picture. These groups of pictures can be collectively put into similar background batches. There are other groups of images, they will be put under general-background-batch, and they will work without the G-picture, as it does not hold any significant advantage. A similar background batch needs to be reworked for better bit encoding for better quality preservation and compression. For this, we can have two types of quantizing methodologies, one is to use the same value as the one used for the general-background-batch (p) and the other is to use another value for the quantization parameter (p'), this is not the same as the one used in the first case and this helps in better rate-distortion values. General-background-batch's valuable frames will be denoted by G_B and we must follow (7) analogy. This means that using the G-picture as a long-term reference, we can quantize any batch of frames with a large valued quantization parameter than the one used for the adjacent frame in the batch of frames.

3.2. Speeding up the algorithm

The foreground units contain the basic coding blocks, which are 4×4 units in size and each input in the coding blocks is classified based on the number of basic blocks present in the foreground. Taking $K(f)$ as input type for basic coding block f and $g_{i,j}$ be a pixel value of basic unit f while for G-picture it is $G_{B_{i,j}}$, then as shown in (9).

$$K(f) = \begin{cases} Y, & \sum_{i=1}^{4} \sum_{j=1}^{4} | g_{i,j} - G_{B_{ij,}}(g)| \leq x \\ H, & \text{otherwise} \end{cases} \quad (9)$$

Here, x is a predefined threshold valued at 80. Then taking in the basic blocks in the group of coding blocks (o is used for its representation), the categories of classes for the coding blocks are calculated with the help of the proportion values of foreground blocks (F_G), its background blocks (B_G) and its hybrid blocks (H_G). The size taken here is $(2N \times 2N)$ computed through (10).

$$\text{Class}(o) = \begin{cases} F_G, & \text{if} 4 \times \left|\left|\{i \middle| K\big(o(i)\big) = H\}\right|\right| / N^2 > \alpha \\ B_G, & \text{if} 4 \times \left|\left|\{i \middle| K\big(o(i)\big) = H\}\right|\right| / N^2 \leq \beta \\ H_G, & \text{if} \alpha \geq 4 \times \left|\left|\{i \middle| K\big(o(i)\big) = H\}\right|\right| / N^2 > \beta \end{cases} \quad (10)$$

$\alpha = 0.5; \quad \beta = 0.0625$

In the traditional HEVC encoder, the encoding value is chosen between $2N \times 2N$ coding blocks or just four recursively-coded parts. To avoid this confusion and reduce time consumption by not calculating and comparing the rate-distortion costs, there is a need for partition termination methods in the HEVC test model. For this, a static background for a large time is used. Each input is considered as a potential coding block and is segregated into the respective blocks as in (10). B_Gs with a value of $N > 8$ will occupy a larger proportion than the other two and that needs an early termination. So, whenever there are 16×16, or $N = 8$ coding blocks then the B_G will be a pure version of the coding blocks and will not undergo further partition. There is also an issue regarding prediction pixels for coding blocks. For better accuracy, it has been decided that only $2N \times 2N$ coding blocks must be used for B_G $N > 8$. The rest will have it for $N \geq 8$ and H_Gs have no asymmetric motions partitions. In addition, the range for searching motion must be at 1 pixel for B_Gs and unchanged for H_Gs and F_Gs.

3.3. Modelling the background and selection

We need to calculate the average of all background frames in a running fashion. J denotes the current frame in training, M is the matrix that has unsigned *-bit integers for average result representation. Then M', that is, the average value is given by (11).

$$M' = (M \times (m - 1) + J + (m >> 1))/m \quad (11)$$

The number of training frames, m, is indicated here. Only one multiply, shift, floor, divide, and three extra operations are performed during this process. The first image, if it is large enough, will be spotted by the algorithm and can be thought of as a large group of frames for minimal time delay in the coding stage. Assume that this batch of frames' HPS has a size that is even. L and $O(X, Y) = 1$ or 0 demonstrates that X and Y have vast amounts of data with different/similar data proportions. Then, $O(J_m)$ of any input picture with thickness m, where m is denoted as $lL + i(l0, i = 0L - 1)$ and his represents the initial image, is determined by (12):

$$O(J_m) = \begin{cases} \text{general} - \text{background} - \text{patch}, & R(J_{l \times L}, G_B) = 1 \\ \text{similar} - \text{background} - \text{patch}, & R(J_{l \times L}, G_B) = 0 \end{cases} \quad (12)$$

for $R(X, Y)$ a 1-pixel range is taken to search in Y the basic units A. This is given by (13). Algorithm 1 mentioned for background modelling.

$$R(X, Y) = \begin{cases} 1, & \text{if } 16 \times ||A(X, Y)/w \times h > 0.8 \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

$$A(X,Y) = \left\{(q,p) \mid \sum_{t,s=1}^{4} \left|X_{4_{4_{q+t}4_{p+s}}} - Y_{4_{q+t}4_{p+s}}\right| \leq 80, p < \frac{w}{4}, q < \frac{h}{4}\right\}$$

Algorithm 1. Algorithm for background modelling

```
Input: Frame $J_m$ of size $h \times w$, where $m = l \times L + i$
Output: $O(J_m)$ type frame with $\text{general} - \text{background} - \text{patch}$ or $\text{similar} - \text{background} - \text{patch}$
if $O(J_m) == O(J_{m-i})$ and $i \neq 0$; return
$X = J_m; Y = G_B;$ $A(X,Y) = \emptyset$
for $q = 1$ to $h/4$ do
    for $p = 1$ to $w/4$ do
        if $\sum_{t,s=1}^{4} \left|X_{4_{4_{q+t}4_{p+s}}} - Y_{4_{q+t}4_{p+s}}\right| \leq 80$
            $A(X,Y) = A(X,Y) \cup \{(q,p)\};$
        end
    end
if $16 \times ||A(X,Y)/w \times h > 0.8$;
    $R(X,Y) = 1;$
else
    $R(X,Y) = 0;$
if $R(X,Y) == 1$ then $O(J_m) = \text{general} - \text{background} - \text{patch}$
    else $O(J_m) = \text{similar} - \text{background} - \text{patch}$
```

In addition, if we take the starting intra image in the low-delay predictor of hierarchy algorithm and quantize it for each similar-background-patch, then the quantization value comes out to be as shown in (14).

$$P_Q(J_{l \times L \times i}) = \begin{cases} P_Q + 1, & \text{if } i = L - 1 \\ P_Q + 2, & \text{if } i = L/2 \\ P_Q + 3, & \text{if } i \neq \frac{L}{2} \text{or } L - 1 \end{cases} \tag{14}$$

Now to effectively calculate the quantization parameters for each general-background-patch frame, we can follow the (15).

$$P_Q(J_{l \times L \times i}) = \begin{cases} P_Q + 2, & \text{if } i = L - 1 \\ P_Q + 4, & \text{if } i \neq L - 1 \end{cases} \tag{15}$$

Next, we must take the G-picture to be quantized at a lesser value for the surrounding frames, as shown in (16).

$$\Delta P_Q = \begin{cases} 5, & \frac{\text{if} D_1}{J_{bp}} > \frac{LS}{3}; \\ 10, & \text{if} \frac{LS}{20} < \frac{D_1}{J_{bp}} < \frac{LS}{3}; \\ 20, & \text{if} \frac{D_1}{J_{bp}} < \frac{LS}{20}; \end{cases} \tag{16}$$

4. EXPERIMENTS AND RESULTS

The entire work has been compared with Wang *et al.* [26] and uses the various saliency detection methods for our evaluation. To maintain uniformity in comparison, we have used the same datasets as mentioned by Wang *et al.* [26]. This will help in evaluating our performance and accuracy in terms of other state the art methods. The collection, referred to as DHF1K [27], has around 1,000 films with a frame rate of 30 and a resolution of 640×360. There are 600 training tests, 300 testing exams, and 100 validation tests. The data from 17 observers is collected using the eye tracker.

4.1. Evaluation metrics employed

The selection of the experiment's evaluation metrics was aided by Bylinskii *et al.* [28]. Area under ROC curve (AUC), Pearson's correlation coefficient (CC), normalized scanpath saliency (NSS). Similarity or histogram intersection (SIM), shuffled AUC, and AUC are the ones that were selected. These measurements have been useful for both self-evaluation and comparison with other cutting-edge techniques for determining video saliency. We have compared our proposed model with these existing models like temporal-spatial feature pyramid network (TSFP-Net) [29], hierarchical decoding for dynamic saliency prediction (HD2S) [30], visual features based convolutional encoder-decoder (ViNet) [31], deep learning approach (DeepVS) for radio frequency (RF)-based vital signs sensing) [32], Chen *et al.* [33], efficient end-to-end audio classification convolutional neural network (ACLNet) [27], spatio-temporal self-attention 3D network (STRA-Net) [34], temporally-aggregating spatial encoder-decoder network (TASED-Net) [35], saliency prediction model with shuffled attentions and correlation (SalSAC) [36], saliency based exponential moving average (SalEMA) [37], unified image and video saliency modeling (UNISAL) [38], and spatial-temporal and self-attention encoding network (STSANet) [26] serves as the foundation for this solution model.

4.2. Results

The comparison among all the mentioned state-of-the-art methods is given in Table 1. As can be discerned from Table 1, the evaluation metrics for the proposed solution have outperformed almost all state-of-the-art methods. This has done best in the SIM metric while ViNet [31] has done best in the sauce. In the remaining list, the performance has been quite good and even the Kullback-Leibler divergence values of the base reference STSANet [26] and the proposed system are 1.344 and 1.297 respectively.

Table 1. Comparison of all the values of the evaluation metrics mentioned for all the state-of-the-art methods along with our proposed system

METHOD	DHF1K				
	CC	NSS	SIM	AUC	sAUC
TSFP-Net [29]	0.517	2.966	0.392	0.912	0.723
HD2S [30]	0.503	2.812	0.406	0.908	0.700
ViNet [31]	0.511	2.872	0.381	0.908	0.729
DeepVS [32]	0.344	1.911	0.256	0.856	0.583
Chen *et al.* [33]	0.476	2.685	0.353	0.900	0.680
ACLNet [27]	0.434	2.354	0.315	0.890	0.601
STRANet [34]	0.458	2.558	0.355	0.895	0.663
TASED-Net [35]	0.470	2.667	0.361	0.895	0.712
SalSAC [36]	0.479	2.673	0.357	0.896	0.697
SalEMA [37]	0.449	2.574	0.466	0.890	0.667
UNISAL [38]	0.490	2.776	0.390	0.901	0.691
STSANet [26]	0.529	3.010	0.383	0.913	0.723
Proposed system	0.547	3.109	0.407	0.933	0.701

This tells us that dissimilarity for our proposed solution is much better and has outperformed once again. The accuracy of the suggested solution is significantly superior than the other evaluated methods, as shown in Figures 1 and 2, because it is much more in line with reality. This proves that the suggested answer is the most accurate and precise of all the alternatives. Figure 1 shows the comparison of all existing models with proposed system. Figure 2 shows the comparison of the ground truths with the proposed system and other state-of-the-art methods.

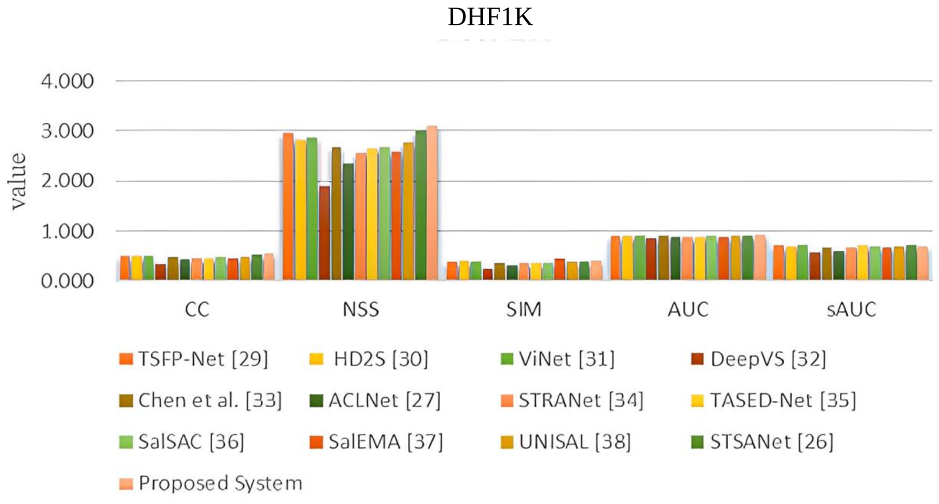

Figure 1. Comparison of methodologies

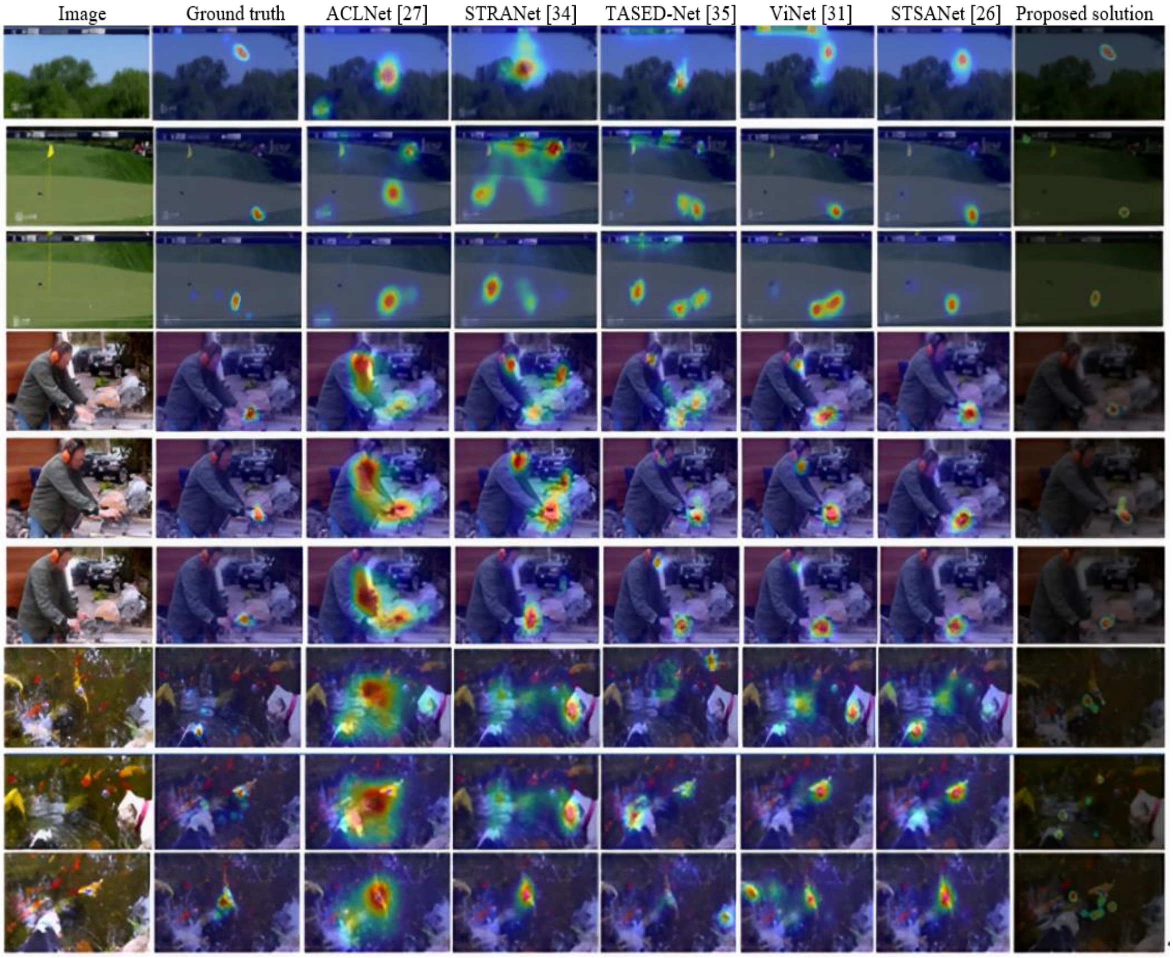

Figure 2. Comparison of the ground truths with the proposed system and other state-of-the-art methods

5. CONCLUSION

In this paper, a modified HEVC technique with spatiotemporal saliency encoding and background adjustment was offered as a potential remedy. The use of the G-picture methodology in the fourth frame as a long-term reference frame is one of two strategies used to make this solution work. Then comes the need to use the coding blocks classification for background segregation for quantization of each frame respectively along with quantization of the G-picture as well. This has led to a reduction in time consumption and coding complexity along with an increase in efficiency and accuracy when the video is compressed. Even though the results display a good increase in almost every evaluation metric chosen for this paper, there is still quite enough room for improvement. We hope that this solution will act as a stepping-stone for other researchers to build on their future solutions in bringing video saliency detection closer to the level of humans and their eye and brain coordination.

REFERENCES

[1] M. Jiang, S. Huang, J. Duan, and Q. Zhao, “SILICON: saliency in context,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 1072-1080, doi: 10.1109/CVPR.2015.7298710.

[2] 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.

[3] 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.

[4] Y. Li, Y. Zhou, L. Xu, X. Yang, and J. Yang, “Incremental sparse saliency detection,” in *2009 16th IEEE International Conference on Image Processing (ICIP)*, Nov. 2009, pp. 3093–3096, doi: 10.1109/ICIP.2009.5414465.

[5] Y. Li, Y. Zhou, J. Yan, Z. Niu, and J. Yang, “Visual saliency based on conditional entropy,” in *ACCV 2009: Computer Vision – ACCV 2009*, 2010, pp. 246–257, doi: 10.1007/978-3-642-12307-8_23.

[6] S. H. Sreedhara, V. Kumar, and S. Salma, “Efficient big data clustering using Adhoc fuzzy C means and auto-encoder CNN,” in *Inventive Computation and Information Technologies*, 2023, pp. 353–368. doi: 10.1007/978-981-19-7402-1_25.

[7] J.-S. Kim, J.-Y. Sim, and C.-S. Kim, "Multiscale saliency detection using random walk with restart," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 2, pp. 198–210, Feb. 2014, doi: 10.1109/TCSVT.2013.2270366.

[8] 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. 2009, pp. 1597–1604. doi: 10.1109/CVPR.2009.5206596.

[9] X. Hou and L. Zhang, "Saliency detection: a spectral residual approach," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2007, pp. 1–8. doi: 10.1109/CVPR.2007.383267.

[10] B. Schauerte and R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction," in *ECCV 2012: Computer Vision – ECCV 2012*, 2012, pp. 116–129, doi: 10.1007/978-3-642-33709-3_9.

[11] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 3166–3173, doi: 10.1109/CVPR.2013.407.

[12] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015, doi: 10.1109/TPAMI.2014.2345401.

[13] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 2814–2821, doi: 10.1109/CVPR.2014.360.

[14] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: contrast based filtering for salient region detection," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 733-740, doi: 10.1109/CVPR.2012.6247743.

[15] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, Jan. 2010, doi: 10.1109/TIP.2009.2030969.

[16] D.-Y. Chen, H.-R. Tyan, D.-Y. Hsiao, S.-W. Shih, and H.-Y. M. Liao, "Dynamic visual saliency modeling based on spatiotemporal analysis," in *2008 IEEE International Conference on Multimedia and Expo*, Jun. 2008, pp. 1085–1088, doi: 10.1109/ICME.2008.4607627.

[17] S.-H. Lee, J.-H. Kim, K. P. Choi, J.-Y. Sim, and C.-S. Kim, "Video saliency detection based on spatiotemporal feature learning," in *2014 IEEE International Conference on Image Processing (ICIP)*, Oct. 2014, pp. 1120–1124, doi: 10.1109/ICIP.2014.7025223.

[18] H. Kim, Y. Kim, J.-Y. Sim, and C.-S. Kim, "Spatiotemporal saliency detection for video sequences based on random walk with restart," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2552–2564, Aug. 2015, doi: 10.1109/TIP.2015.2425544.

[19] C.-W. Lin and Y.-R. Lee, "Fast algorithms for DCT-domain video transcoding," in *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)*, 2001, vol. 1, pp. 421–424, doi: 10.1109/ICIP.2001.959043.

[20] J. Zhang, S. Li, and C.-C. J. Kuo, "Compressed-domain video retargeting," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 797–809, Feb. 2014, doi: 10.1109/TIP.2013.2294541.

[21] O. Sukmarg and K. R. Rao, "Fast object detection and segmentation in MPEG compressed domain," in *2000 TENCON Proceedings. Intelligent Systems and Technologies for the New Millennium (Cat. No.00CH37119)*, 2000, vol. 2, pp. 364–368, doi: 10.1109/TENCON.2000.892290.

[22] P. Liu and K. Jia, "Low-complexity saliency detection algorithm for fast perceptual video coding," *The Scientific World Journal*, vol. 2013, pp. 1–15, 2013, doi: 10.1155/2013/293681.

[23] S. H. Khatoonabadi, I. V. Bajić, and Y. Shan, "Compressed-domain correlates of fixations in video," in *Proceedings of the 1st International Workshop on Perception Inspired Video Processing*, Nov. 2014, pp. 3–8, doi: 10.1145/2662996.2663008.

[24] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 3888–3901, Sep. 2012, doi: 10.1109/TIP.2012.2199126.

[25] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, and C.-W. Lin, "A video saliency detection model in compressed domain," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 1, pp. 27–38, Jan. 2014, doi: 10.1109/TCSVT.2013.2273613.

[26] Z. Wang *et al.*, "Spatio-temporal self-attention network for video saliency prediction," *IEEE Transactions on Multimedia*, vol. 25, pp. 1161–1174, 2023, doi: 10.1109/TMM.2021.3139743.

[27] 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.

[28] 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.

[29] Q. Chang and S. Zhu, "Temporal-spatial feature pyramid for video saliency detection," *Prepr arXiv210504213*, 2021.

[30] G. Bellitto, F. P. Salanitri, S. Palazzo, F. Rundo, D. Giordano, and C. Spampinato, "Hierarchical domain-adapted feature learning for video saliency prediction," *International Journal of Computer Vision*, vol. 129, no. 12, pp. 3216–3232, Dec. 2021, doi: 10.1007/s11263-021-01519-y.

[31] S. Jain, P. Yarlagadda, S. Jyoti, S. Karthik, R. Subramanian, and V. Gandhi, "ViNet: pushing the limits of visual modality for audio-visual saliency prediction," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2021, pp. 3520–3527, doi: 10.1109/IROS51168.2021.9635989.

[32] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "DeepVS: a deep learning based video saliency prediction approach," in *ECCV 2018: Computer Vision – ECCV 2018*, 2018, pp. 625–642, doi: 10.1007/978-3-030-01264-9_37.

[33] J. Chen, H. Song, K. Zhang, B. Liu, and Q. Liu, "Video saliency prediction using enhanced spatiotemporal alignment network," *Pattern Recognition*, vol. 109, Jan. 2021, doi: 10.1016/j.patcog.2020.107615.

[34] Q. Lai, W. Wang, H. Sun, and J. Shen, "Video saliency prediction using spatiotemporal residual attentive networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1113–1126, 2020, doi: 10.1109/TIP.2019.2936112.

[35] K. Min and J. Corso, "TASED-Net: temporally-aggregating spatial encoder-decoder network for video saliency detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 2394–2403, doi: 10.1109/ICCV.2019.00248.

[36] X. Wu, Z. Wu, J. Zhang, L. Ju, and S. Wang, "SalSAC: a video saliency prediction model with shuffled attentions and correlation-based ConvLSTM," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12410–12417, Apr. 2020, doi: 10.1609/aaai.v34i07.6927.

[37] P. Linardos, E. Mohedano, J. J. Nieto, N. E. O'Connor, X. Giro-i-Nieto, and K. McGuinness, "Simple vs complex temporal recurrences for video saliency prediction," *Prepr arXiv190701869*, 2019.

[38] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," in *ECCV 2020: Computer Vision – ECCV 2020*, 2020, pp. 419–435, doi: 10.1007/978-3-030-58558-7_25.

[39] 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*, pp. 1–1, 2020, doi: 10.1109/TCSVT.2019.2911396.

[40] 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.

[41] 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.

[42] 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.

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

[44] M. Banitalebi-Dehkordi, A. Ebrahimi-Moghadam, M. Khademi, and H. Hadizadeh, “No-reference quality assessment of HEVC video streams based on visual memory modelling,” *Journal of Visual Communication and Image Representation*, vol. 75, Feb. 2021, doi: 10.1016/j.jvcir.2020.103011.

[45] L. Jiang, M. Xu, Z. Wang, and L. Sigal, “DeepVS2.0: a saliency-structured deep learning method for predicting dynamic visual attention,” *International Journal of Computer Vision*, vol. 129, no. 1, pp. 203–224, Jan. 2021, doi: 10.1007/s11263-020-01371-6.

[46] X. Zhang, L. Liang, Q. Huang, Y. Liu, T. Huang, and W. Gao, “An efficient coding scheme for surveillance videos captured by stationary cameras,” in *Proceedings of SPIE - The International Society for Optical Engineering*, Jul. 2010, doi: 10.1117/12.863522.

[47] M. Paul, W. Lin, C.-T. Lau, and B.-S. Lee, “Explore and model better I-frames for video coding,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 9, pp. 1242–1254, Sep. 2011, doi: 10.1109/TCSVT.2011.2138750.

[48] X. Zhang, T. Huang, Y. Tian, and W. Gao, “Background-modeling-based adaptive prediction for surveillance video coding,” *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 769–784, Feb. 2014, doi: 10.1109/TIP.2013.2294549.

[49] K. Tian, Z. Lu, Q. Liao, and N. Wang, “Video saliency detection based on robust seeds generation and spatio-temporal propagation,” in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Oct. 2017, pp. 1–6, doi: 10.1109/CISP-BMEI.2017.8301936.

[50] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, “Overview of the high-efficiency video coding (HEVC) standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012, doi: 10.1109/TCSVT.2012.2221191.

BIOGRAPHIES OF AUTHORS

Sharada P. Narasimha received the B.E. degree from Bapuji Institute of Engineering and Technology, Belagavi, India, in 2003, M.Tech. degree in computer science and engineering from SJC Institute of Technology, India, in 2013. She is pursuing towards her Ph.D. in VTU-RC at Sri Venkateswara College of Engineering. 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, computer networks, IoT, MEMS, and embedded systems. She can be contacted at this email: sharada_p2k22@redffimail.com.

Dr. Sanjeev C. Lingareddy received his Ph.D. in the year of 2012 from Jawaharlal Nehru Technological Universit, 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 this email: sclingareddy@gmail.com.