Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

Video Saliency Detection Using Modified Hevc And Background Modelling

Mrs. Sharada P N[1] , Dr. S C Lingareddy[2]

[1]Research scholar VTU.

[2]Prof SVCE Bangalore.

ABSTRACT

The world is now heavily dependent on live streaming and video conferences with the advent of the Covid – 19 pandemic. This change has led to the formation of several algorithms that specialise in video saliency. Our proposed algorithm is a modified HEVC algorithm that employs background modelling and implication of classification into coding blocks. With the help of G-Picture in the fourth long term reference position and usage of coding blocks, the overall coding complexity along with time consumption has been reduced as well as the efficiency of compression has increased. The algorithm is tested on the DHF1K dataset and has been compared with several state-of-the-art methods and the final result showcases that the proposed solution has the best overall compression efficiency and accuracy.

A. INTRODUCTION

The world has been trying to copy the process of the human eye of filtering the unnecessary parts of what we are viewing and only considering the important ones. Image saliency has been a well-researched field while video saliency has not been so. [1] The SALICON dataset has been a major stepping stone in the formulation of better image saliencies. [3] Itti et al. and [2] Milanfar and Seo have been the basis of research in this field. [4] and [5] uses combinational blocks randomly based of the neighbouring blocks. [6] involves graphs for saliency and [7] has a similar approach but it also includes random walk models to obtain colour, motion, compact and intensity features into a graph for stationary distribution of Markov chain as a saliency map. [8] uses random walk with restart methodology that helps in motion distinctiveness and correction of abrupt change and temporal consistency.

[9]'s working model involves discriminant centre surround hypothesis which combines colour, spatial and temporal saliencies with help of temporal gradients. [10] has used feature extraction to partially decode data and then uses local and global spatiotemporal features to form a map an then is fused with motion vectors and DCT coefficients to get the final result.

[11] has improved the DCT domain transcoder into the DDT for faster extraction for partial low frequency coefficients. [12] has used the low-level compression feature from the bitstream. [13] has employed object recognition for speed boosting and [14-17] has used color clustering and region merging that is based on spatiotemporal similarities, classification based on region and pixel edge extraction for video saliency.

[31-34] have a lot of very useful saliency methods. One of them has G-Picture usage at second reference frame to reduce HEVC complexity, other uses quantization parameters along with Background Reference Prediction (BRP) and Background Difference Prediction (BDP). Even small coding blocks called Coding Units were introduced to lower complexity and increase compression efficiency. All these works have helped in formulating our solution and have been a great help to this project.

This paper proposes a novel model of modified HEVC algorithm that background modelling using hierarchy prediction structure. It has two components, the first being the modification of G-Picture usage as a reference frame, making it the fourth reference frame and not the second as mentioned in other research papers, and quantizing it with a relatively smaller valued parameter along with the inclusion of coding blocks for reduced complexity. The second component is the segregation of each coding block into F_G, B_G and H_G. Each of these components is sped up differently on the basis of the data available in the G-Picture. there is an inclusion of another modification where the coding block portioning is terminated early so as to avoid extra coding and computation.

This paper contains five sections in total. The first section handles the introduction, the second maintains a record of the related works for this paper. The third explains the proposed system's mathematical and coding aspect while fourth showcases the results of the experiments conducted with the dataset DHF1K. Then, the fifth section concludes the paper.

B. LITERATURE SURVEY

This section will give a brief look into the various researches and experiments that have helped us in building up our solution. Starting with [35] we have an improved version of the HEVC algorithm in which the perception redundancy has been reduced for better compression value. It involves combination of motion estimation and each block during compression phase. Then the convolutional neural network uses spatiotemporal method to obtain the final saliency map.

[36] and [37] are survey papers that deal with the differences that make the saliency methods not as efficient as the human eye-brain coordination. It concludes with the fact that coding complexity needs a reduction for betterment of the conventional saliency algorithms.

[38] is the dataset that this paper has used for its experimentation. It is called the Dynamic Human Fixation 1K (or DHF1K) that predicts fixations during dynamic scene viewing. With one thousand great quality varied video sequences from 17 observers using an eye tracker. This has also proposed a state-of-the-art video saliency method named ACL Net (Attentive CNN-LSTM Network. It has also compared its results with other methods with different datasets, namely Hollywood-2 and UCF Sports. It was one of the fastest methods proposed till now.

[39] has used super-saliency and has employed manual algorithmic annotations of smooth pursuits for saliency fixation and training slicing convolutional neural networks. These results are formulated using the help of 26 publicly available dynamic saliency models.

[40] is research that proposes a 3-dimensional convolutional encoder-decoder architecture for prediction in dynamic scenes. There are two subnetworks in the encoder that extracts the spatial and temporal features of each frame and undergoes intermediate fusion. Then the decoder enlarges the features in spatial dimensions and aggregates temporal information. It is trained in an end-to-end manner and also experimented upon the DHF1K dataset. [41] is a deep neural network-based video saliency prediction method and is named as DeepVS2.0. it has helped in comparing our results and check how well we have performed with respect to other state-of-the-art methods. It uses object-to-motion convolutional neural network (OM-CNN) that uses spatiotemporal features to form the intra-frame saliency map.

[18] is the base reference on the basis of which we shall compare the results of our work to evaluate our performance. It is called Spatio-Temporal Self-Assessment (STSA Net) ad uses layers of 3-dimensional convolutional backbone for the mapping of the saliency.

C. PROPOSED SYSTEM

a. Optimizing Low-Delay HPS efficiently

We shall calculate the rate distortion cost C as $C = \mu\eta + B$, where η denotes number of bits and μ denotes Lagrange multiplier. If Q as quality of reconstructed video and $\Psi(I_i, p)$ be the rate distortion cost of encoding the $i-th$ image I_i where m will be the input frames with p denoting the quantization parameter of the coding units with cost function τ, we get $C = \sum_i^m \Psi(I_i, p) = \sum_i^m \sum_r \tau(p, I_{i,r}, Q_{i,r,p}\, U_{i,r,p})$. Here $U_{i,r,p}$ represents the motion vectors and $Q_{i,r,p}$ is the data prediction quantized with p.

If we take a smaller quantization value as p', it provides better reference for a images $(I_{j+1} \sim I_{j+a})$. We then get a better costing equation $C' = T_1 + T_2 + T_3 + T_4 + T_5$ where $T_1 = \sum_{i=1}^{j-1} \Psi(I_i, p)$, $T_2 = \Psi(I_i, p')$, $T_3 = \sum_{i=j+1}^{j+a} \sum_{l=1}^{Q_i} \tau(p, I_{i,t(i,l)}, Q_{i,t(i,l),p,}\, U_{i,t(i,l),p})$, $T_4 = \sum_{i=j+1}^{j+a} \sum_{r=1}^{n_i} \tau(p, I_{i,e(i,l)}, Q_{i,e(i,l),p'}, U_{i,e(i,l),p'})$ and $T_5 = \sum_{i=j+a+1}^{m} \Psi(I_i, p)$. If we change T_4 to costing I_j and T_5 is cost for I_{j+a}. The modified costing equation using p comes out to be $C = T_1 + X + T_3 + Y + T_5$ where $X = \Psi(I_j, p)$ and $Y = \sum_{i=j+1}^{j+a} \sum_{r=1}^{n_i} \tau(p, I_{i,e(i,l)}, Q_{i,e(i,l),p,}\, U_{i,e(i,l),p})$. Then taking difference of the two types of costing rates comes out to be $C - C' = (Y - T_4) - (T_2 - X)$. By making T_4's quantization value lesser we solve the inequality $Y - T_4 = \sum_{i=j+1}^{j+a} \sum_{r=1}^{n_i} (\tau(p, I_{i,e(i,l)}, Q_{i,e(i,l),p,}\, U_{i,t(i,l),p}) - \tau(p, I_{i,e(i,l)}, Q_{i,e(i,l),p'}, U_{i,e(i,l),p'})) > 0$.

b. Speeding up the algorithm

The foreground units are in 4×4 size and 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 $G_{B_{i,j}}$, we get

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

Here, x is a predefined threshold valued at 80. Now taking basic coding blocks o, of size $(2N \times 2N)$), the categories of classes for the coding blocks is calculated with the help of the proportion values of foreground blocks (F_G), its background blocks (B_G) and its hybrid blocks (H_G).

$$Class(o) = \begin{cases} F_G, & if4 \times \left|\left|\{i|K(o(i)) = H\}\right|\right| /N^2 > \alpha \\ B_G, & if4 \times \left|\left|\{i|K(o(i)) = H\}\right|\right| /N^2 \le \beta \\ H_G, & if\alpha \ge 4 \times \left|\left|\{i|K(o(i)) = H\}\right|\right| /N^2 > \beta \end{cases} \quad (2)$$

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

c. Modelling the Background and selection

J denotes current frame in training, M is the matrix that has unsigned *-bit integers for average result representation. Then $'$, that is, the average value, is given by $M' = (M \times (m-1) + J + (m >> 1))/m$. Here m denotes number of training frames. Suppose that the hierarchal prediction structure for this batch of frames has an even valued size L and $O(X, Y) = 1$ or 0 showcases that X and Y have dissimilar/similar proportions of data in large quantity. Then $O(J_m)$ of any input image of m thinness where m is written as $l \times L + i(l \ge 0, i = 0 \sim L-1)$ which also means that $m = l \times L$ and his represents the first image and this is calculated by

$$O(J_m) = \begin{cases} general - background - patch, & R(J_{l\times L}, G_B) = 1 \\ similar - background - patch, & R(J_{l\times L}, G_B) = 0 \end{cases} \quad (3)$$

For $R(X, Y)$ a 1-pixel range is taken to search in Y the basic units A. This is given by

$$R(X, Y) = \begin{cases} 1, & \text{if } 16 \times ||A(X, Y)/w \times h > 0.8 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

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

The quantization value for a low delay predictor of the hierarchy algorithm, for each similar background patch, comes out to be

(5)

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

Then the effective calculation comes out to be

$$P_Q(J_{l\times L\times i}) = \begin{cases} P_Q + 2, & \text{if } i = L - 1 \\ P_Q + 4, & \text{if } i \neq L - 1 \end{cases} \quad (6)$$

Next, we must take the G-Picture to be quantized at a lesser value with respect to the surrounding frames,

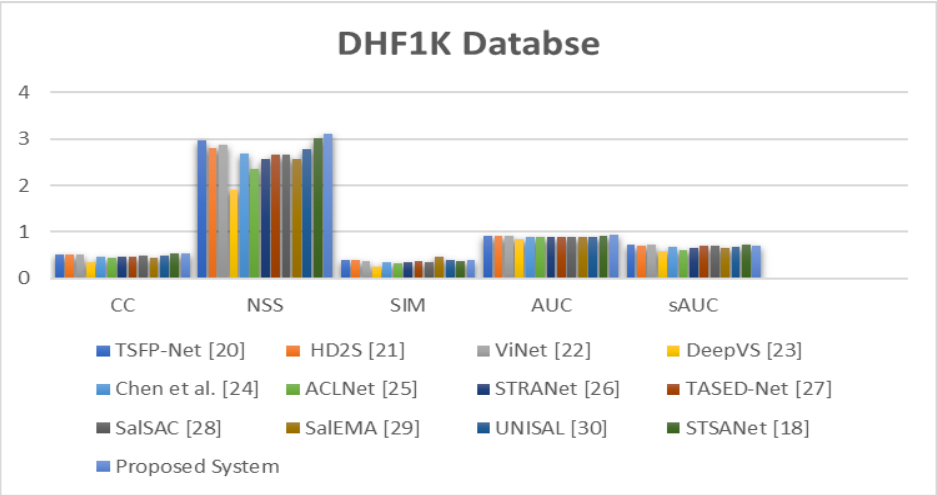
$$\Delta P_Q = \begin{cases} 5, & \text{if} \frac{D_1}{J_{bp}} > \frac{LS}{3}; \\ 10, & \text{if} \frac{LS}{20} < \frac{D_1}{J_{bp}} < \frac{LS}{3}; \\ 20, & \text{if} \frac{D_1}{J_{bp}} < \frac{LS}{20}; \end{cases} \quad (7)$$

D. EXPERIMENTS AND RESULTS

This paper has been compared with STSA Net [18] as a base reference for evaluation of performance. The dataset used is same as the base paper mentioned to pertain uniformity. Its name is DHF1K and has a huge library of videos wit 30fps settings and 640×360 resolution. There are 100 validations, 300 testing and 600 training tests. The eye tracker is used to obtain data from 17 observers.

The evaluation metrics have been chosen from [19] and they are Pearson's Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS), Similarity or histogram intersection (SIM), Area Under ROC Curve (AUC) and Shuffled AUC (s AUC). The other state-of-art methods for video saliency are TSFP-Net [20], HD2S [21], Vi Net [22], Deep VS [23], Chen et al. [24], ACL Net [25], STRA-Net [26], TASED-Net [27], Sal SAC [28], Sal EMA [29] and UNISAL [30].

The comparison among all the mentioned state-of-the-art-methods is given in Graph 1.



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

As it can be discerned from the graph above, the evaluation metrics for the proposed solution has outperformed almost all state-of-the-art methods. The Sal EMA [29] has done best in the SIM metric while Vi Net [22] has done best in s AUC. Other than the mentioned methodologies, our proposed solution has the best values in almost all metrics.

Figure 1 helps in seeing how the accuracy of the proposed solution is much better than the other compared methods as it is much closer to the ground truth. This is a testament that the accuracy and precision of the proposed solution is the best amongst all the other methods. Figure 1 is given below.

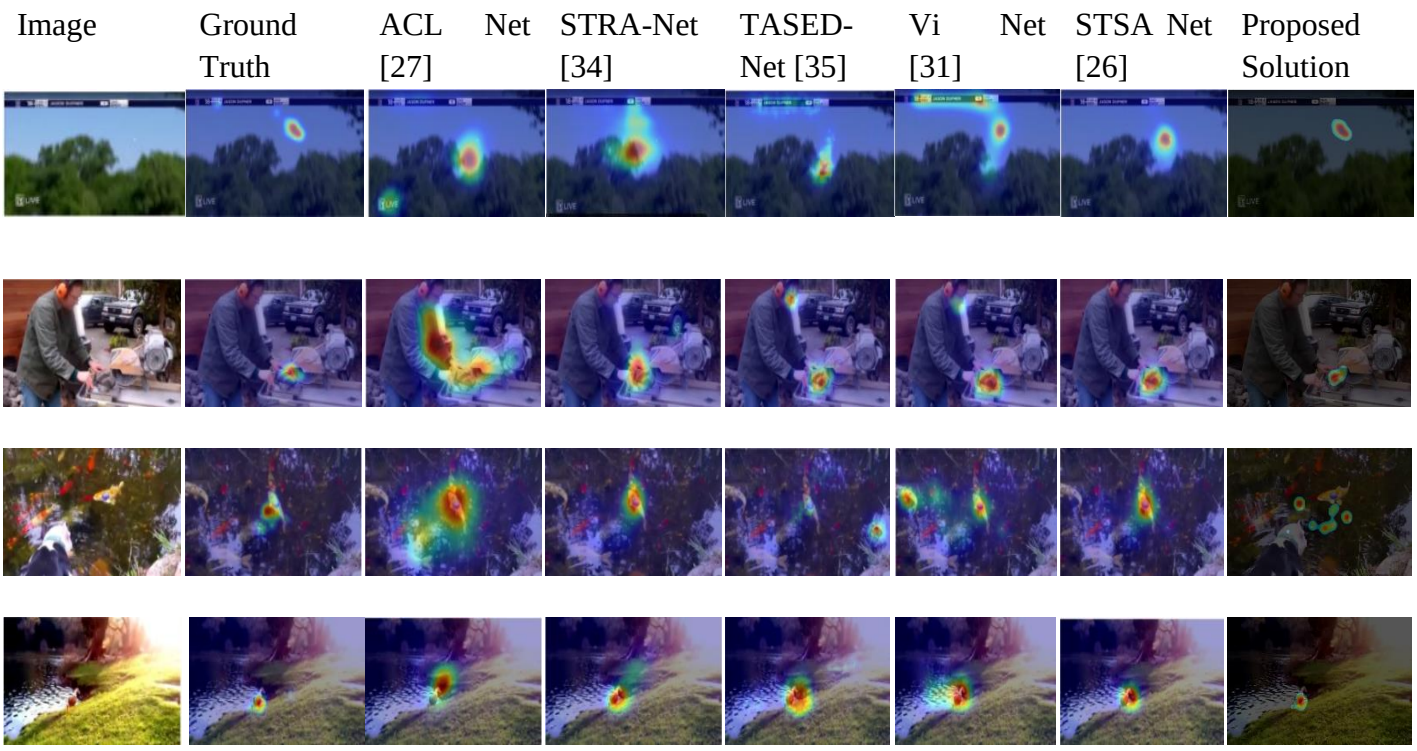


Figure 1.: Comparison among the ground truths with the proposed system and other state-of-the-art-methods.

E. CONCLUSION

The world has seen several image saliency researches but not much has been done in the field of video saliency. This paper proposes a novel solution which is a modified HEVC algorithm that employs spatiotemporal saliency along with background modification. The G-Picture is set to be the fourth long-term reference frame and then using coding blocks for classification of background segregation along with quantization of each frame takes place. This helps in reduction of coding complexity, time consumption and overall increase in compression efficiency. The results show that in all the evaluation metrics chosen for this paper, the proposed algorithm has performed the best I overall evaluation in comparison to other state-of-the-art saliency methodologies.

REFERENCES

1. M. Jiang, S. Huang, J. Duan, and Q. Zhao. SALICON: Saliency in context. In CVPR, 2015.
2. L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.
3. H. J. Seo, P. Milanfar, Static and space-time visual saliency detection by self-resemblance, J. Vision 9 (12) (2009) 1–27.
4. Y. Li, Y. Zhou, L. Xu, X. Yang, J. Yang, Incremental sparse saliency detection, in: Proceedings of the IEEE International Conference on Image Processing, 2009, pp. 3093–3096.
5. Y. Li, Y. Zhou, J. Yan, J. Yang, Visual saliency based on conditional entropy, in: Proceedings of the Asian Conference on Computer Vision, 2009, pp. 246–257.
6. J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in: Proceedings of the Advances in Neural Information Processing Systems, 2006, pp. 545–552.
7. J.-S. Kim, J.-Y. Sim, C.-S. Kim, Multiscale saliency detection using random walk with restart, IEEE Trans. Circuits Syst. Video Technol. 24 (2) (2014) 198–210.
8. H. Kim, Y. Kim, J.-Y. Sim, C.-S. Kim, Spatiotemporal saliency detection for video sequences based on random walk with restart, IEEE Trans. Image Process. 24 (8) (2015) 2552–2564.
9. [D.-Y. Chen, H.-R. Tyan, D.-Y. Hsiao, S.-W. Shih, H.-Y. Liao, Dynamic visual saliency modeling based on spatiotemporal analysis, in: Proceedings of the IEEE Multimedia and Expo Conference, 2008, pp.1085–1088.
10. S.-H. Lee, J.-H. Kim, K. P. Choi, J.-Y. Sim, C.-S. Kim, Video saliency detection based on spatiotemporal feature learning, in: Proceedings of the IEEE International Conference on Image Processing, 2014, pp. 1120–1124.
11. C.-W. Lin, Y.-R. Lee, Fast algorithms for DCT-domain video transcoding, in: IEEE Trans. Image Process., Vol. 1, 2001, pp. 421–424.
12. J. Zhang, S. Li, C.-C. Kuo, Compressed-domain video retargeting, IEEE Trans. Image Process. 23 (2)(2014) 797–809.
13. O. Sukmarg, K. Rao, Fast object detection and segmentation in MPEG compressed domain, in: Proceedings of the IEEE TENCON, Vol. 3, 2000, pp. 364–368.

14. P. Liu, K. Jia, Low-complexity saliency detection algorithm for fast perceptual video coding, The Scientific World Journal 2013.
15. S. H. Khatoonabadi, I. V. Baji´, Y. Shan, Compressed-domain correlates of fixations in video, in: Proceedings of the 1st International Workshop on Perception Inspired Video Processing, 2014, pp. 3–8.
16. Y. Fang, Z. Chen, W. Lin, C.-W. Lin, Saliency detection in the compressed domain for adaptive image retargeting, IEEE Trans. Image Process. 21 (9) (2012) 3888–3901.
17. Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, C.-W. Lin, A video saliency detection model in compressed domain, IEEE Trans. Circuits Syst. Video Technol. 24 (1) (2014) 27–38.
18. Z. Wang et al., "Spatio-Temporal Self-Attention Network for Video Saliency Prediction," in IEEE Transactions on Multimedia, doi: 10.1109/TMM.2021.3139743.
19. Z. Bylinskii, T. Judd, A. Oliva, A. Torralba and F. Durand, "What Do Different Evaluation Metrics Tell Us About Saliency Models?," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 3, pp. 740-757, 1 March 2019, doi: 10.1109/TPAMI.2018.2815601.
20. Q. Chang, S. Zhu, and L. Zhu, “Temporal-spatial feature pyramid for video saliency detection,” ar Xiv reprint arXiv:2105.04213, 2021.
21. 2] G. Bellitto, F. P. Salanitri, S. Palazzo, F. Rundo, D. Giordano, and C. Spampinato, “Hierarchical domain-adapted feature learning for video saliency prediction,” arXiv reprint arXiv:2010.01220, 2020
22. [23] 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,” ar Xiv reprint arXiv:2012.06170, 2020.
23. L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, “Deep VS: A deep learning based video saliency prediction approach,” in European Conference on Computer Vision (ECCV), 2018, pp. 625–642.
24. J. Chen, H. Song, K. Zhang, B. Liu, and Q. Liu, “Video saliency prediction using enhanced spatiotemporal alignment network,” Pattern Recognition, vol. 109, p. 107615, 2021.
25. 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.
26. 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.
27. K. Min and J. J. Corso, “TASED-Net: Temporally-aggregating spatial encoder-decoder network for video saliency detection,” in IEEE international Conference on Computer Vision (ICCV), 2019, pp. 2394–2403.
28. X. Wu, Z. Wu, J. Zhang, L. Ju, and S. Wang, “Sal SAC: A video saliency prediction model with shuffled attentions and correlation-based convlstm,” in AAAI Conference on artificial intelligence, 2020, pp. 12 410–12 417.

29. 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," in British achine Vision Conference (BMVC), 2019, pp. 1–12.
30. R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," in European Conference on Computer Vision (ECCV), 2020, pp. 419–435.
31. X. Zhang, L. Liang, Q. Huang, and W. Gao, "An efficient coding scheme for surveillance videos captured by stationary cameras," in Proc. Vis. Commun. Image Process., Jul. 2010, pp. 1–10.
32. M. Paul, W. Lin, C.-T. Lau, and B.-S. Lee, "Explore and model better I-frames for video coding," IEEE Trans. Circuits Syst. Video Technol., vol. 21, no. 9, pp. 1242–1254, Sep. 2011.
33. X. Zhang, T. Huang, Y. Tian and W. Gao, "Background-Modeling-Based Adaptive Prediction for Surveillance Video Coding," in IEEE Transactions on Image Processing, vol. 23, no. 2, pp. 769-784, Feb. 2014, doi: 10.1109/TIP.2013.2294549.
34. G. J. Sullivan, J. Ohm, W. Han and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1649-1668, Dec. 2012, doi: 10.1109/TCSVT.2012.2221191
35. 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," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 7, pp. 1946-1959, July 2020, doi: 10.1109/TCSVT.2019.2911396.
36. Man Hoang, Trinh & Zhou, Jinjia. (2021). Recent Trending on Learning based Video Compression: A survey. Cognitive Robotics. 1. 10.1016/j.cogr.2021.08.003.
37. Borji, Ali. (2019). Saliency Prediction in the Deep Learning Era: Successes and Limitations. IEEE transactions on pattern analysis and machine intelligence. PP. 10.1109/TPAMI.2019.2935715.
38. W. Wang, J. Shen, J. Xie, M. -M. Cheng, H. Ling and A. Borji, "Revisiting Video Saliency Prediction in the Deep Learning Era," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 220-237, 1 Jan. 2021, doi: 10.1109/TPAMI.2019.2924417.
39. Startsev, Mikhail & Dorr, Michael. (2019). Supersaliency: A Novel Pipeline for Predicting Smooth Pursuit-Based Attention Improves Generalisability of Video Saliency. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2961835.
40. H. Li, F. Qi and G. Shi, "A Novel Spatio-Temporal 3D Convolutional Encoder-Decoder Network for Dynamic Saliency Prediction," in IEEE Access, vol. 9, pp. 36328-36341, 2021, doi: 10.1109/ACCESS.2021.3063372.
41. Jiang, Lai & Xu, Mai & Wang, Zulin & Sigal, Leonid. (2021). DeepVS2.0: A Saliency-Structured Deep Learning Method for Predicting Dynamic Visual Attention. International Journal of Computer Vision. 129. 10.1007/s11263-020-01371-6.