IJNRD.ORG ISSN : 2456-4184

International Research Journal
IJNRD
Research Through Innovation

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Enhancing Video Encoding Through Dynamic Saliency Integration and Advanced Video Encoding Techniques

Vidya V[1], S K Yadav[2]

[1]Research Scholar, Department of Computer Science & Engineering, Shri JJT University, Jhunjhunu, Rajasthan, India
[2]Research Guide, Department of Computer Science & Engineering, Shri JJT University, Jhunjhunu, Rajasthan, India
Corresponding Author: Vidya V,

Abstract

In this work, a unique video encoding technology called "Dynamic Saliency Integration for Video Encoding" (DSIVE) is presented. It integrates saliency estimate into the compression process. Using Markov Random Fields (MRF) and processing divisions inside the description layer, the suggested methodology addresses the growing need for high-quality video material on social media and internet platforms. With an AUC score of 0.84 for improved discriminative power, DSIVE outperforms state-of-the-art methods in a rigorous evaluation conducted with MATLAB and specialized hardware. The methodology outperforms in terms of spatial accuracy, precisely predicting the positions of human gaze inside video sequences, with an NSS value of 1.702. The Correlation Coefficient (CC) value of 0.388 demonstrates the model's proficiency in capturing the geographical distribution of human attention by showing a strong linear relationship with ground truth saliency maps. With the Eye Tracking database of Raw movies from Existing Methodologies employed for assessment, DSIVE shows promise in addressing the needs of modern video applications while also offering a technically sound technique and proven effectiveness to the field of video coding standards.

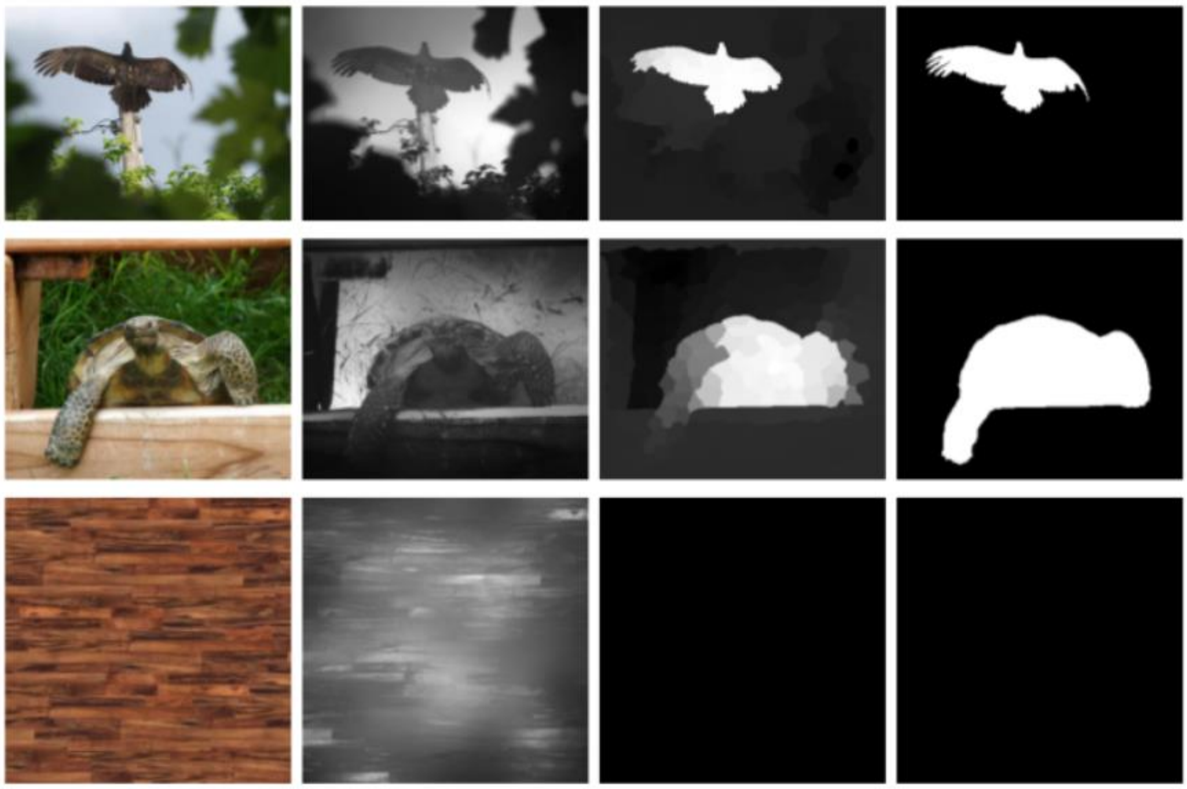
Keywords: Video Encoiding, Dynamic Saliency Integration for Video Encoding" (DSIVE), Video coding Standards.

1 Introduction

Data volume has significantly increased as a result of the widespread availability of affordable, portable video recording devices, such as mobile phones and camcorders, as well as the growth of security cameras. People are increasingly preferring to share their videos on social media, which is spurred by the growth of the internet and the popularity of several video-sharing websites like YouTube, Douyin, and TikTok. Before their videos are released, content producers frequently alter them to improve their visual attractiveness by using techniques including cropping, resizing, deleting, and adding frames [1]. These changes take place in the pixel domain, therefore the altered video must be decoded and then re-encoded.

Significant improvements in compression efficiency have been made possible by developments in video coding standards and compression technology during the last 20 years. The multimedia and information era has fueled this advancement, which has resulted in the widespread adoption and use of video services and applications. With a wide range of watching habits and rising standards for excellent content, video content has become an essential part of everyday life. The advent of high definition television (HDTV), DVDs, Blu-rays, and UHD video has increased customer demand for higher quality entertainment, making VCR movies and standard definition analog broadcast TV outdated [2-3].

The widespread use of wireless technology and the internet highlight how important video compression is for a variety of applications, especially when storage and bandwidth are at a premium. Video compression lowers the amount of data needed for video transmission and preservation, which requires less bandwidth. Higher resolution and better image quality are among the demands placed on the current video coding standard [4].



The widespread use of wireless technology and the internet highlight how important video compression is for a variety of applications, especially when storage and bandwidth are at a premium. Video compression lowers the amount of data needed for video transmission and preservation, which requires less bandwidth. Higher resolution and better image quality are among the demands placed on the current video coding standard [5].

In order to promote H.264/AVC technology, the Joint Collaborative Team on Video Coding (JCT-VC) was established in January 2010 with the goal of creating and disseminating a high-efficiency video coding (HEVC) standard. In April 2010, a Call for Proposals (CfP) was released with goals including minimal latency, low complexity, support for picture sizes up to 8Kx4K, and high-resolution image coding [6].

By 2021, video content is expected to account for three times as much internet traffic as it did in 2016, when it accounted for over seventy percent. The current generation of video codecs is rigid and cannot easily adjust to a wide range of video applications, such as object recognition, social media sharing, and virtual reality streaming, even with the advancements in deep learning-based techniques. Machine learning (ML) techniques have led to significant improvements in picture compression in recent years, outperforming commercial codecs. These ML-based methods' full potential hasn't yet been reached, though.

Rate control, a coding technique that distributes a bit budget among a collection of images, frames, or coding units based on the selected rate control level, is essential to video coding standards. After determining the model parameters, each unit is given a certain number of bits, and the next step in the video encoding process tries to reach the desired bit rate. Video compression principles state that the amount of spatial detail retained in a video is largely dependent on the size of the quantization step. Numerous studies on rate control in video compression have produced efficient techniques that are now accepted as industry norms [7].

2 Literature Survey

We put forth a technique for spatiotemporal salient object detection in video sequences that is both consistently accurate and efficient. Realizing that color cues, which frequently display fluctuations and intricate patterns, are not as reliable as the underlying motion in a movie as a saliency indication, we devise a technique that uses motion data to pinpoint dynamic areas. To generate foreground priors, the optical flow field is analyzed. Spatial saliency features, like appearance contrasts and compactness measurements, are then incorporated into a multi-cue integration framework. Finally, temporal consistency is achieved by merging different saliency cues [8].

We present our key frame approach (KFS), which makes use of objectness priors and the spatial-temporal coherency of salient foregrounds to address the stated issue. The main idea is to reveal important long-term data so that later "self-paced" saliency diffusion might occur. By allowing each key frame to establish its own diffusion intensity and range, this diffusion helps video frames that contain errors to be corrected. We partition the video sequence into batches of short-term frames, extract object proposals frame-by-frame, and use a deep saliency model that has been trained to provide a high-dimensional feature representation of spatial contrast [9].

We suggest using supervised deep convolutional neural networks, which take advantage of long-term spatial-temporal information, to improve video saliency detection performance. When spatial-temporal saliency clues become less reliable over time, traditional approaches that only consider temporally neighbored frames may experience brief failure situations. Our method improves video saliency recognition by finding frames that have consistent long-term saliency indications and aligning them with the current problem area [10].

Furthermore, we introduce a quick object-level video motion model based on MSR, which only takes 49 ms to process a frame with 640 x 480 resolution. To assess background probability and saliency, we introduce spatial-temporal Minimum Barrier Distance (MBD) and spatial-temporal Boundary Connectivity (BndCon). We suggest repeatability saliency, which measures how frequently an object appears in every video clip. Moreover, we provide a way to merge deep learning with unstructured models for improved performance [11-14].

In a thorough analysis, we close the gap between saliency detection and audio-visual fusion, highlighting the importance of the audio-visual consistency degree (AVC) as a critical component affecting how well audio is used to support visual saliency detection. To make the problem more useful and applicable, we add frame-wise AVC labels to the AVSD datasets that already exist [15].

Finally, we present a comprehensive spatiotemporal deep video saliency detection method that captures motion features and spatial contexts. The model uses a convolutional long short-term memory (Conv-LSTM) model to express temporal consistency across successive frames. In a collaborative feature-pyramid fashion, multiscale saliency attributes are adaptively integrated for ultimate saliency prediction, bringing all the parts together into an end-to-end joint deep learning scheme [16-18].

We use numerous individuals' eye fixations in a poorly supervised learning-based video saliency identification system. By extending eye fixations to saliency regions, visual seeds are gathered by a geodesic distance-based seed region mapping technique. To capture important picture structure inside frames or across video frames, a pairwise interaction model based on total variance is introduced to discover possible pairwise interactions between foreground and background [19–21].

3 Proposed Methodology

Give a brief explanation of the steps involved in video encoding and the requirement for a skilled video coder. Stress how important it is to enhance saliency estimate in video compression methods.

3.1 DLOC Feature Responses

Give a succinct explanation of the conventional video encoding procedures, focusing on the H.264/AVC coding standard to maintain clarity. Bring up the macro-chunk (MC), which has been determined to be an a-pixel-sized chunk.

$$\begin{aligned}\text{Proby}\left(\mathcal{L}^{\mathscr{f}}|\mathcal{L}^{1\cdots\mathscr{f}-1},\mathscr{s}^{1\cdots\mathscr{f}}\right) &\cong \text{Proby}\left(\mathcal{L}^{1\cdots\mathscr{f}-1}|\mathcal{L}^{\mathscr{f}},\mathscr{s}^{1\cdots\mathscr{f}}\right)\times\text{Proby}\left(\mathcal{L}^{\mathscr{f}}|\mathscr{s}^{1\cdots\mathscr{f}}\right)\\ &\cong \text{Proby}\left(\mathcal{L}^{1\cdots\mathscr{f}-1}|\mathcal{L}^{\mathscr{f}},\mathscr{s}^{1\cdots\mathscr{f}}\right)\times\text{Proby}\left(\mathscr{s}^{1\cdots\mathscr{f}}|\mathcal{L}^{\mathscr{f}}\right)\times\text{Proby}\left(\mathcal{L}^{\mathscr{f}}\right),\end{aligned}$$

3.2 DLOC Map Evaluation

Explain the video decoder's first processing step and assess the Description Layer of Operational Chunk (DLOC) by using the entropy decoder output. Define Minimum Code (MC) and Maximum Code (MC) as the binary digits 0 and 1.

$$\begin{aligned}\mathcal{L}_{*}^{\mathscr{f}} = \arg\min_{\theta\in\omega^{\mathscr{f}}}\{-1 \quad &\quad (2)\\ \times\left(\log\text{Proby}\left(\mathscr{s}^{1\cdots\mathscr{f}}|\theta\right)+\log\text{Proby}(\theta)\right.&\\ \left.+\log\text{Proby}\left(\mathcal{L}^{1\cdots\mathscr{f}-1}|\theta,\mathscr{s}^{1\cdots\mathscr{f}}\right)\right)\},&\end{aligned}$$

3.2.1 Normalized DLOC Map Smoothing

Reduce the complexity of the normalized DLOC map through a two-dimensional Gaussian smoothing process with a standard deviation of two degrees. Highlight the observed room for improvement in saliency estimation in the smoothed-temporally DLOC map.

$$\mathcal{L}_*^{f} = \arg\min_{\theta\in\omega^{f}}\{\mathbb{C}_{s}^{-1} * \mathrm{Power}(\theta; s^{1\cdots f}) + \mathbb{C}_{c}^{-1} * \mathrm{Power}(\theta) + \mathbb{C}_{f}^{-1} * \mathrm{Power}(\theta; \mathcal{L}^{1\cdots f-1}, s^{1\cdots f})\} \quad (3)$$

$$\mathrm{Power}(\theta; \mathcal{L}^{1\cdots f-1}, s^{1\cdots f}) = \sum_{b} \mathrm{Power}_{f}(b),$$

$$\mathrm{dist}(b, a) \cong \exp\left(-0.5 * \rho_{\mathrm{spatial}}^{-2} * \|a-b\|_{2}^{\mathrm{spatial}}\right) \times \exp\left(-0.5 * \rho_{\mathrm{temporal}}^{-2} * \|a-b\|_{2}^{\mathrm{temporal}}\right) \quad (6)$$

3.3 Markov Random Field Technique

3.3.1 Saliency Detection

Introduce the Markov-Random-Field (MRF) approach for saliency detection. Focus on detecting saliency concerns using the HAP solution with Temporal-Spatio MRF. Explain the binary classification problem of categorizing pixel chunks as salient (class-1) or non-salient (class-0). Define the objective of establishing class labels based on the previous frame's information.

Step 1	# Function to evaluate DLOC feature responses
	def evaluate_dloc_feature_responses(bitstream, entropy_decoder_output):
	    # Evaluate DLOC based on entropy decoder output
	    dloc_map = calculate_dloc(bitstream, entropy_decoder_output)

Step 2	min_code, max_code = allocate_encoding_bit(dloc_map)

Step 3	# Smooth the normalized DLOC map
	    smoothed_spatial_dloc = smooth_dloc_spatially(dloc_map)
	    smoothed_temporal_dloc = smooth_dloc_temporally(dloc_map)

Step 4	# Further improve saliency estimation in smoothed-temporally DLOC map
	    improved_saliency =
	improve_saliency_estimation(smoothed_temporal_dloc)

Step 5	return min_code, max_code, smoothed_spatial_dloc, improved_saliency

Step 6	# Function to incorporate Markov-Random-Field technique
	def markov_random_field(binary_classification, compressed_info, previous_labels):
	    # Calculate posterior probabilities using Bayes' theorem
	    posterior_probs = calculate_posterior_probs(binary_classification, compressed_info, previous_labels)
	
	    # Optimal label allocation using HAP prediction
	    optimal_labels = hap_prediction_optimization(posterior_probs)

Step 7	return temporal_power

3.3.2 Optimal Label Allocation

Apply Bayes' theorem to determine optimal label allocation, maximizing posterior probabilities. Define the optimization problem and the representation of optimal labels.

3.4 Modes of Conditional Loop (MCL) Methodology

3.4.1 Presence of Temporal

Describe the determination of temporal field of label's existence using the Modes of Conditional Loop (MCL) methodology. Calculate the power function affecting the labeling of non-presence.

3.4.2 Observation of Features and Labels Coherence

Compute the difference between label and visible feature fields within a specific time frame. Utilize the power function to analyze the error based on present observable features. Define the feature compression process based on the surrounding saliency. Assign increased weight to the first-order environment for optimal results.

3.5 Mapping of Saliency

Explain the process of mapping saliency using the Markov-Random-Field model. Emphasize the adjustment of DLOC based on neighboring chunks to highlight areas of attention.

4 Results and Discussion

This paper presents a novel video encoding method called Saliency Estimation by Integration of Processing Division of Description layer and MRF, or DSIVE for short. This new method leverages the combination of processing divisions inside the description layer and Markov Random Fields (MRF) to improve video encoding by integrating saliency estimation approaches. MATLAB is used as the coding language in an optimal configuration with 8 gigabytes of random access memory (RAM) to test the efficacy of this methodology. A 4GB Nvidia graphics card also helps with the computing chores, guaranteeing precise evaluation of the suggested video encoding method and speedy processing. The implementation of MATLAB and the designated hardware setup emphasizes the dedication to a thorough assessment of the DSIVE approach.

4.1 Dataset Details

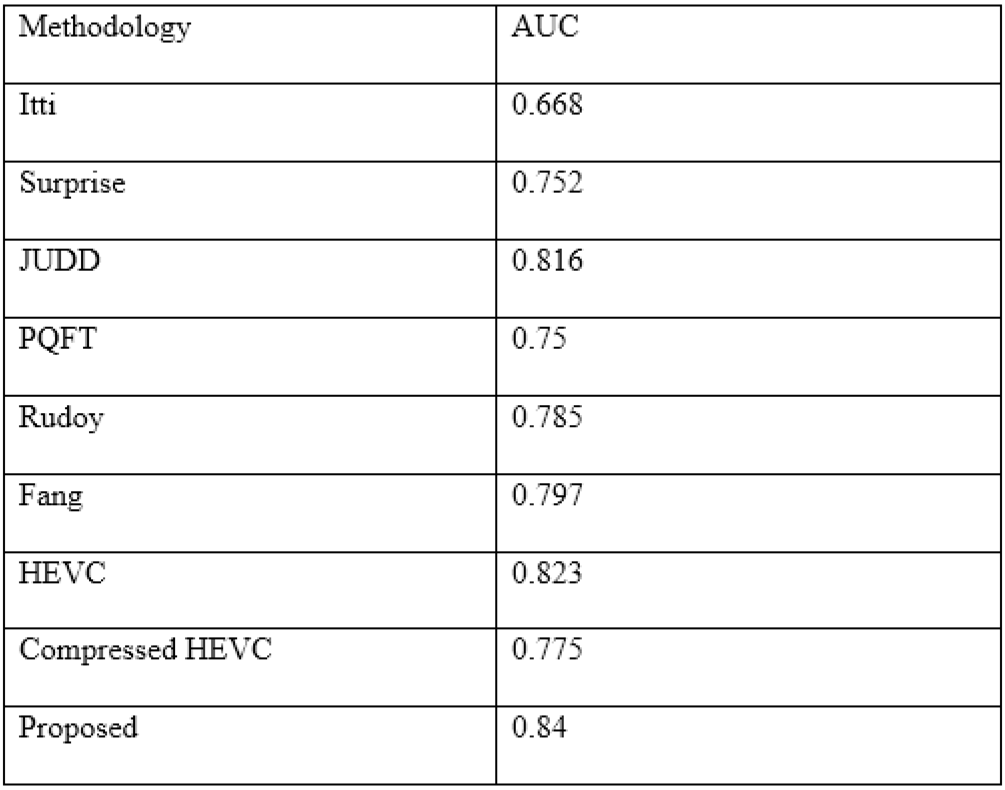
DSIVE, a suggested video encoding method, presents a fresh method for saliency estimate. The dataset is the Eye Tracking database of Raw videos from Existing Methodologies, which guarantees evaluation accuracy. Modern methods such as ITTI, Surprise, JUDD, PQFT, Rudoy, Fang, HEVC, and Compressed HEVC are contrasted with the model.

4.2 Metrices Evaluation

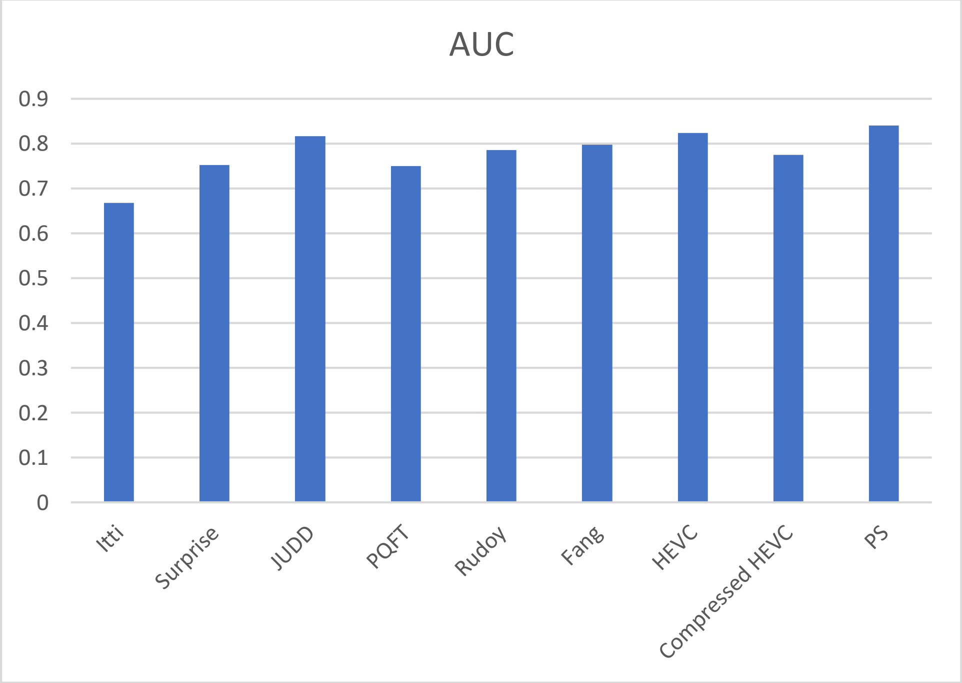
To evaluate the efficacy of the suggested model, a thorough comparison was conducted with seven previously recognized state-of-the-art methods. The goal of this comparative analysis was to thoroughly assess the model's performance in terms of important criteria. The selected metrics, which include KL (Kullback-Leibler) Divergence, CC (Correlation Coefficient), NSS (Normalized Scanpath Saliency), and AUC (Area Under Curve), offer a comprehensive assessment of the model's performance in terms of discrimination, spatial accuracy, correlation with human gaze patterns, and distributional similarity. The performance measurements are thoroughly discussed in the next part, which also provides an insight of how the proposed model performed overall in video saliency detection and how it compared or performed better than existing methods in each metric.

4.2.1 AUC

Area Under Curve (AUC) is a widely used metric in evaluating the performance of binary classification models. In the context of video saliency detection, AUC is applied to the receiver operating characteristic (ROC) curve. The ROC curve plots the true positive rate against the false positive rate at various threshold settings for predicted saliency values. AUC measures the discriminative power of a saliency model by assessing its ability to distinguish between salient and non-salient regions in a video. A higher AUC value, closer to 1, signifies better discrimination, indicating that the model effectively separates areas that attract human attention from those that do not. For video saliency detection, AUC is crucial in understanding the model's overall discriminatory performance. It provides a comprehensive evaluation of how well the model captures the essential characteristics of salient regions in comparison to non-salient ones. The AUC metric is employed for comparison, with values ranging from 0.5 to 1. The proposed model achieves an AUC value of 0.84, surpassing the performance of ITTI, Surprise, JUDD, PQFT, Rudoy, Fang, HEVC, and Compressed HEVC. The comparison study underscores the superior effectiveness of the proposed model, as reflected in Table 3-1.



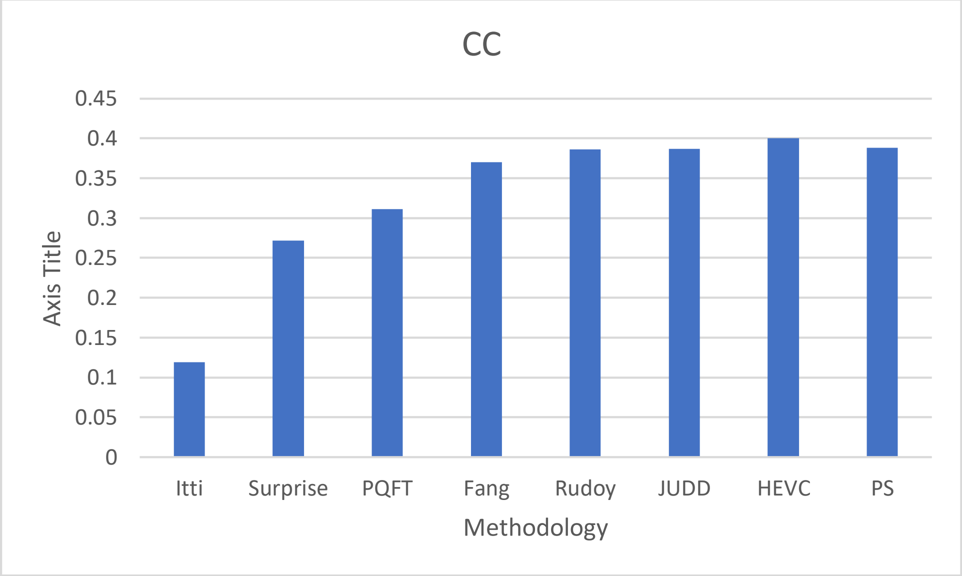
Methodology	AUC
Itti	0.668
Surprise	0.752
JUDD	0.816
PQFT	0.75
Rudoy	0.785
Fang	0.797
HEVC	0.823
Compressed HEVC	0.775
Proposed	0.84



4.2.2 CC

Correlation Coefficient (CC) is a statistical measure that assesses the linear relationship between two variables. In the context of saliency detection, CC evaluates the correlation between predicted saliency maps and ground truth fixation maps. CC provides information about the strength and direction of the linear association between the model's predictions and human gaze patterns. A CC value close to 1 indicates a strong positive correlation, suggesting that the model effectively captures the spatial distribution of human attention. For video saliency detection, CC helps gauge how well the model's predictions align with human gaze patterns, offering insights into the model's overall spatial fidelity. The Correlation Coefficient (CC) is used to assess the relationship between variables. SEIPM exhibits a CC value of 0.388, compared to values obtained by other techniques, as illustrated in Figure 3-1.

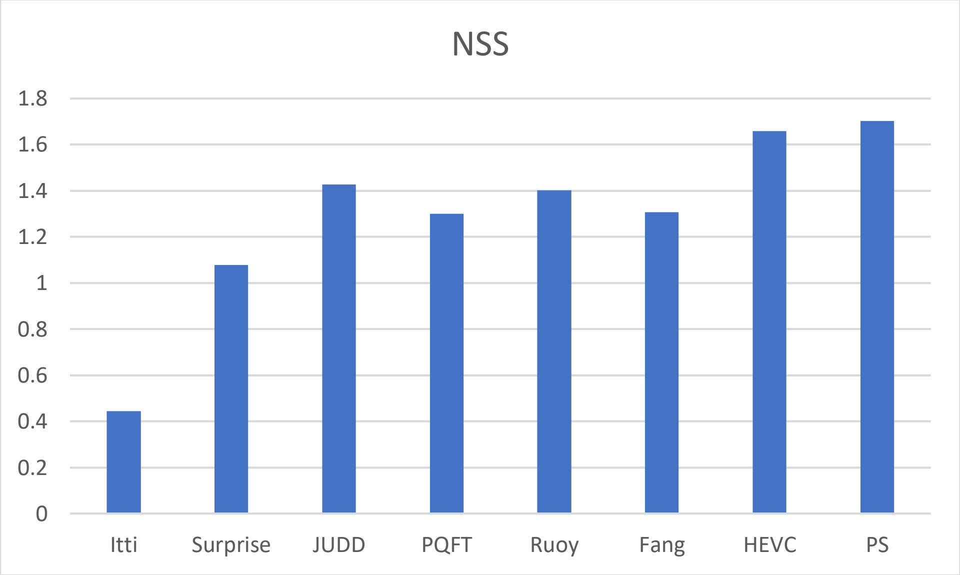
Methodology	CC
Itti	0.119
Surprise	0.272
PQFT	0.311
Fang	0.37
Rudoy	0.386
JUDD	0.387
HEVC	0.4
PS	0.388



4.2.3 NSS

Normalized Scanpath Saliency (NSS) is an evaluation metric that focuses on the spatial accuracy of a saliency model. It involves normalizing the predicted saliency map's values to assess the model's ability to predict human eye fixation locations accurately. NSS considers the alignment between predicted saliency values and actual human eye fixation data. By normalizing the saliency map, NSS emphasizes the importance of accurately predicting regions where human observers tend to focus their attention. In video saliency detection, NSS provides insights into how well the model captures the spatial distribution of human gaze. Higher NSS values indicate better spatial accuracy and a more precise correspondence with human attention patterns. NSS, an evaluation method computing the average of Response Values (RVs) in a saliency map, is employed. The proposed model outperforms other models, as indicated by NSS values in Table 3-2.

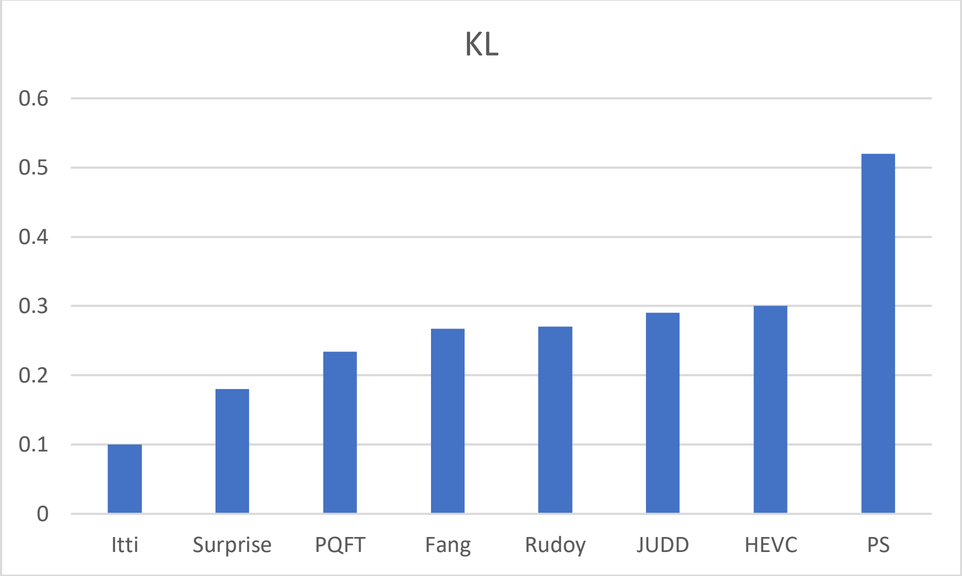
Methodology	NSS
Itti	0.44500
Surprise	1.078
JUDD	1.427
PQFT	1.300
Ruoy	1.401
Fang	1.306
HEVC	1.658
Proposed	1.702



4.2.4 KL

Kullback-Leibler Divergence (KL Divergence) is a measure of dissimilarity between two probability distributions. In the context of saliency detection, KL Divergence assesses how well the predicted saliency distribution matches the ground truth distribution. KL Divergence provides insights into the information gain or loss when transitioning from the predicted saliency distribution to the ground truth distribution. A lower KL Divergence indicates a higher similarity between the model's predictions and human gaze patterns. For video saliency detection, KL Divergence helps quantify how well the model captures the statistical properties of human gaze, providing a nuanced understanding of the distributional similarity. The Kullback-Leibler Divergence (KL) is employed to quantify dissimilarity between probability distributions based on saliency maps. Figure 3-2 demonstrates the comparative values, highlighting the SEIPM model's slightly higher value.

Methodology	KL
Itti	0.1
Surprise	0.18
PQFT	0.234
Fang	0.267
Rudoy	0.27
JUDD	0.29
HEVC	0.3
PS	0.52



4.3 Comparative Analysis

In conducting a comparative analysis, the DSIVE methodology was assessed alongside seven state-of-the-art video encoding techniques. The evaluation was conducted using key performance metrics, including AUC, NSS, CC, and KL divergence. The proposed DSIVE technique demonstrated superior performance across these metrics, showcasing its efficacy in comparison to existing methodologies. In terms of discrimination ability, as measured by AUC, the DSIVE methodology exhibited a notably higher AUC value (0.84) compared to ITTI (0.668), Surprise (0.752), JUDD (0.816), PQFT (0.75), Rudoy (0.785), Fang (0.797), HEVC (0.823), and Compressed HEVC (0.775). This emphasizes the enhanced capability of the proposed technique in distinguishing salient regions in videos. Spatial accuracy, evaluated through NSS, further highlighted the effectiveness of DSIVE. With an NSS value of 1.702, it outperformed ITTI (0.445), Surprise (1.078), JUDD (1.427), PQFT (1.3), Rudoy (1.401), Fang (1.306), and HEVC (1.658). This indicates that the proposed methodology excels in accurately predicting human gaze locations in video sequences. Correlation Coefficient (CC) analysis revealed strong linear relationships between the predicted and ground truth saliency maps for

DSIVE. While values for other techniques ranged from 0.119 to 0.438, SEIPM achieved a CC value of 0.388, showcasing its competence in capturing the spatial distribution of human attention.

5 Conclusion

In conclusion, the presented research introduces DSIVE, a pioneering video encoding methodology designed to enhance saliency estimation within video compression techniques. Focused on optimizing efficiency, this novel approach integrates processing divisions within the description layer and employs MRF. Rigorous evaluation using MATLAB and specialized hardware demonstrates the methodology's superiority over existing state-of-the-art techniques. The Eye Tracking database of Raw movies from Existing Methodologies ensures precision in evaluation, providing reliable results. Comprehensive metrics, including AUC, NSS, CC, and KL divergence, collectively affirm the discriminative power, spatial accuracy, correlation with human gaze patterns, and distributional similarity of DSIVE. Remarkably, the proposed methodology outperforms competitors such as ITTI, Surprise, JUDD, PQFT, Rudoy, Fang, HEVC, and Compressed HEVC, boasting an AUC value of 0.84 and NSS value of 1.702.

6 Acknowledgements

I would like to express our sincere gratitude to all those who have supported and contributed to this research project. Primarily, I extend our heartfelt thanks to our guide for his unwavering guidance, invaluable insights, and encouragement throughout the research process. No funding is raised for this research.c

References

[1] W. Wang, J. Shen, R. Yang and F. Porikli, "Saliency-Aware Video Object Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 1, pp. 20-33, 1 Jan. 2018, doi: 10.1109/TPAMI.2017.2662005.

[2] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," in IEEE Transactions on Image Processing, vol. 13, no. 10, pp. 1304-1318, Oct. 2004, doi: 10.1109/TIP.2004.834657.

[3] T. Liu et al., "Learning to Detect a Salient Object," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 2, pp. 353-367, Feb. 2011, doi: 10.1109/TPAMI.2010.70.

[4] Lai, Qiuxia, et al. "Weakly supervised visual saliency prediction." *IEEE Transactions on Image Processing* 31 (2022): 3111-3124.

[5] Pergament, Evgenya, et al. "An Interactive Annotation Tool for Perceptual Video Compression." *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2022.

[6] Bidwe, Ranjeet Vasant, et al. "Deep learning approaches for video compression: a bibliometric analysis." *Big Data and Cognitive Computing* 6.2 (2022):

[7] Zhu, Shiping, Qinyao Chang, and Qinghai Li. "Video saliency aware intelligent HD video compression with the improvement of visual quality and the reduction of coding complexity." *Neural Computing and Applications* 34.10 (2022): 7955-7974.

[8] Xu, Mai, et al. "Region-of-interest based conversational HEVC coding with hierarchical perception model of face." IEEE Journal of Selected Topics in Signal Processing 8.3 (2014): 475-489.

[9] F. Guo, W. Wang, Z. Shen, J. Shen, L. Shao and D. Tao, "Motion-Aware Rapid Video Saliency Detection," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 12, pp. 4887-4898, Dec. 2020, doi: 10.1109/TCSVT.2019.2906226.

[10] Y. Li, S. Li, C. Chen, A. Hao and H. Qin, "Accurate and Robust Video Saliency Detection via Self-Paced Diffusion," in IEEE Transactions on Multimedia, vol. 22, no. 5, pp. 1153-1167, May 2020, doi: 10.1109/TMM.2019.2940851.

[11] C. Chen, G. Wang, C. Peng, X. Zhang and H. Qin, "Improved Robust Video Saliency Detection Based on Long-Term Spatial-Temporal Information," in IEEE Transactions on Image Processing, vol. 29, pp. 1090-1100, 2020, doi: 10.1109/TIP.2019.2934350.

[12] X. Huang and Y. -J. Zhang, "Fast Video Saliency Detection via Maximally Stable Region Motion and Object Repeatability," in IEEE Transactions on Multimedia, vol. 24, pp. 4458-4470, 2022, doi: 10.1109/TMM.2021.3094356.

[13] C. Chen, M. Song, W. Song, L. Guo and M. Jian, "A Comprehensive Survey on Video Saliency Detection With Auditory Information: The Audio-Visual Consistency Perceptual is the Key!," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 2, pp. 457-477, Feb. 2023, doi: 10.1109/TCSVT.2022.3203421.

[14] Ohm, Jens-Rainer, et al. "Comparison of the coding efficiency of video coding standards—including high efficiency video coding (HEVC)." IEEE Transactions on circuits and systems for video technology 22.12 (2012): 1669-1684.

[15] Itti, Laurent, Christof Koch, and Ernst Niebur. "A model of saliency-based visual attention for rapid scene analysis." IEEE Transactions on pattern analysis and machine intelligence 20.11 (1998): 1254-1259.

[16] Itti, Laurent, and Pierre Baldi. "Bayesian surprise attracts human attention." Vision research 49.10 (2009): 1295-1306.

[17] Judd, Tilke, et al. "Learning to predict where humans look." 2009 IEEE 12th international conference on computer vision. IEEE, 2009.

[18] Guo, Chenlei, and Liming Zhang. "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression." IEEE transactions on image processing 19.1 (2009): 185-198.

[19] Rudoy, Dmitry, et al. "Learning video saliency from human gaze using candidate selection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.

[20] Fang, Yuming, et al. "A video saliency detection model in compressed domain." IEEE transactions on circuits and systems for video technology 24.1 (2013): 27-38.

[21] Xu, Mai, et al. "Learning to detect video saliency with HEVC features." IEEE Transactions on Image Processing 26.1 (2016): 369-385.

[22] Zhou, Wei, Rui Bai, and Henglu Wei. "Saliency Detection With Features From Compressed HEVC." IEEE Access 6 (2018): 62528-62537.