IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Dynamic Saliency Modeling for Enhanced Video Analysis: A Robust Spatio-Temporal Approach (RSTA)

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

This research introduces a dynamic saliency modeling system for video analysis, combining long-term inter-batch information and color contrast computation to enhance robustness. Leveraging background and foreground appearance models, RGB history handling, and precise saliency adjustment, the proposed method ensures accuracy in capturing visual patterns over time. Spatio-temporal gradient mapping and Robust Principal Component Analysis contribute to contrast-based saliency mapping, demonstrating superior performance in compression efficiency and precision. Experimental results, evaluated against state-of-the-art algorithms, showcase the system's consistent excellence across diverse video sequences and resolutions. The proposed algorithm Robust Spatio-Temporal Approach (RSTA) stands as a promising contribution to computer vision and video analysis applications, affirming its potential impact on future research in the field.

Keywords: Saliency detection, Video analysis, Contrast based saliency

1 Introduction

In the realm of computer vision, researchers have long sought to replicate the sophisticated mechanisms of the human eye and brain. The human brain's ability to discern essential details from visual scenes has been a driving force behind numerous studies, particularly in the domain of image saliency. While models like Itti's have made substantial contributions to understanding image saliency through techniques like Fourier transformation and frequency tuning, the transition from image to video saliency introduces new complexities [1].

Video saliency detection is uniquely challenging due to the dynamic nature of video frames, leading to increased demands on memory allocation and computational processing. Previous methodologies have attempted various approaches, such as computing space-time-saliency maps, motion saliency maps, or fusing static and dynamic saliency mapping. Unfortunately, these methods often accumulate errors, resulting in frequent incorrect detections [2].

To refer these issues, this paper introduces a modified spatio-temporal fusion saliency detection method. The proposed approach begins by creating a spatiotemporal background to enhance saliency around

foreground objects. Subsequent adjustments are made to general saliency strategies, specifically targeting improvements in both motion and color saliencies. A key innovation lies in the incorporation of cross-frame super pixels and a one-to-one spatial-temporal fusion strategy [3]. These enhancements aim to elevate the overall accuracy and precision of the saliency detection process, particularly during compression.

In essence, this refers presents innovative perspective on video SD (saliency detection), delving into the technical nuances of the proposed solution [4]. By offering a detailed exploration of the methodology, the paper contributes to the ongoing discourse on advancing video saliency detection techniques.

Video saliency detection is a crucial task in computer vision, aiming to emulate the intricate mechanisms of human visual attention within dynamic visual content. Unlike image saliency, where identifying conspicuous elements within a single static frame is the primary goal, video saliency detection involves analyzing a sequence of frames over time [5]. The human visual system naturally prioritizes certain regions or objects in a scene based on factors like motion, color, and contrast. Mimicking this process in machines requires the development of sophisticated algorithms that can discern the salient regions in a video, providing valuable insights for applications such as video summarization, object recognition, and action prediction.

The challenges in video saliency detection arise from the temporal dimension advanced by the dynamic nature of videos. Every frame in a video contributes to the overall understanding of the scene, and the salient regions may change over time. Consequently, effective video saliency models must consider not only spatial features but also temporal dynamics. The temporal aspect introduces complexities related to memory allocation, computational processing, and the need for robust algorithms capable of handling diverse content [6].

One prominent avenue of research in video saliency detection involves the fusion of spatial and temporal information. The spatio-temporal contrast in a video is a critical factor, and models often leverage gradient maps to capture changes in both color and motion across frames. This comprehensive understanding of spatio-temporal dynamics allows the algorithm to identify regions that stand out over time, contributing to the overall saliency map. Additionally, methods that integrate spatial information across frames, such as linear iterative clustering and boundary-aware smoothing, play a pivotal role in refining the saliency detection process [7].

Matrix decomposition techniques, particularly those employing RPCA, have gained prominence in video saliency detection. These techniques aim to decompose the video data into sparse and low-level components, allowing for the correction of detections and enhancing the accuracy of super-pixel correspondences. By creating feature subspaces for color and motion, these techniques align super pixels for the infusion of saliency mapping, providing a nuanced analysis of video content [8].

The integration of machine learning approaches, such as CNNs and LSTM networks, has further advanced video saliency detection. These networks can learn intricate patterns and dependencies in video sequences, improving the accuracy along with efficiency of saliency prediction. Attention mechanisms within these networks enable the model to focus on specific regions, akin to the human visual system's selective attention.

Evaluation of video saliency detection algorithms is crucial for benchmarking their performance. Researchers commonly employ metrics such as Area Under the Curve (AUC), Similarity (SIM), Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS), and Kullback-Leibler Divergence (KL) to quantitatively assess the accuracy along with robustness of their models. These metrics provide a

comprehensive evaluation, considering aspects like how well the model predicts human gaze patterns and the consistency of saliency maps across different video sequences [9].

Recent research in video saliency detection has focused on addressing specific challenges, such as handling complex scenes, improving computational efficiency, and refining the accuracy of saliency maps. The proposed solutions often draw inspiration from cognitive models of human vision, aiming to enhance the interpretability and reliability of video saliency predictions. As video content continues to proliferate across various domains, from surveillance to entertainment, the development of robust and efficient video saliency detection algorithms remains a vibrant area of research, with implications for numerous real-world applications.

2 Related Work

This section provides insights into influential research papers that significantly contributed to the development of the proposed algorithm. A comprehensive survey, documented in, explores diverse video saliency methodologies, delineating their respective advantages and disadvantages. Additionally, mirrors this structure while delving into the challenges inherent in replicating human eye-brain coordination and proposing strategies to overcome these hurdles.

Stands out with valuable contributions, presenting the Dynamic Human Fixation 1K (DHF1K) database for identifying fixations during dynamic scene viewing. The Attentive CNN-LSTM Network (ACLNet) within augments the original CNN-LSTM model, facilitating rapid end-to-end saliency learning. Refines smooth pursuits (SP) logic through manual annotations, incorporating fixation points and SP salient locations via trained Slicing Convolutional Neural Networks [9].

Within the realm of High-Definition Video Compression, transforms HEVC algorithms using a spatial saliency algorithm based on motion vectors, leading to enhanced compression efficiency. Introduces a salient object segmentation approach employing conditional random field (CRF) and saliency measure, incorporating statistical frameworks and local features [10-11]. Utilizes spatiotemporal fusion with uncertainty in statistics to gauge visual saliency, while employs the geodesic robustness methodology to derive saliency maps [12].

Significantly contributes to our solution by leveraging super-pixels and adaptive color quantization [13]. Its methodology measures the difference between spatial distance and histograms, crucial for obtaining super-pixel saliency maps. provides a comprehensive overview of evaluation metrics, guiding the paper's methodology. The literature review concludes with Section 5 summarizing the paper's content, with Section 3 and 4 detailing the proposed algorithm, methodologies, modifications, and experimental comparisons.

3 Proposed Methodology

3.1 Dynamic Saliency Modeling for Robust Adjustment

to enhance robustness in the proposed system by leveraging a combination of long-term inter-batch information and color contrast computation. This is achieved through the introduction of background (B_M) and foreground (F_M) appearance models, both represented by matrices. The RGB history of each super pixel in all regions is meticulously handled using specific equations. Saliency adjustment is a key focus, involving the calculation of both inter and intra-saliency values, where color and motion saliency are intricately integrated. To refine these saliency values temporally and spatially, a smoothing process is employed, ensuring a more accurate and stable representation of visual patterns over time.

Step 1	**Color and Motion Salience Calculation** *if visual_data_has_color_and_motion:* *extract_rgb_and_motion_gradients()*
Step 2	**Saliency Adjustment** *if background_and_foreground_models_available:* *calculate_saliency_adjustment()*
Step 3	**Contrast-Based Saliency Mapping** *if video_sequence_divided_into_groups:* *for each frame F_k in G_i:* *modify_frame_using_linear_iterative_clustering(F_k*
Step 4	**Accuracy Boosting** *if matrix_M_provided:* *decompose_matrix_M()*
Step 5	**Foreground Rough Estimation** *if need_to_reduce_incorrect_detections:* *locate_rough_foreground()*
Step 6	**Pixel-Based Saliency Mapping Infusion** *if color_and_motion_saliency_maps_available:* *integrate_saliency_maps()*
Step 7	**Accuracy Enhancement** *if need_for_accuracy_enhancement:* *apply_robust_principal_component_analysis()*
Step 8	**Temporal Batch Adjustment** *if inner_temporal_batch_adjustment_required:* *diffuse_inner_temporal_batch()*
Step 9	**Final Output Calculation** *if all_steps_executed:* *calculate_final_output()*

3.2 Spatio-Temporal Contrast Mapping

video sequence effectively by dividing it into concise groups and applying modifications to individual frames using linear iterative clustering coupled with boundary-aware smoothing. To capture the dynamics of the visual content, the methodology involves the calculation of spatio-temporal gradient maps, integrating both color and motion gradients. Motion contrast of the i-th super pixel is computed through a defined equation, while color saliency is determined similarly, considering RGB notations. Spatial information integration is a critical step in this process, where both motion contrast and color saliency undergo

smoothing to refine their values. This comprehensive approach ensures a meticulous analysis of the video frames, considering both motion and color aspects for an enriched understanding of the visual content.

$$MC_i = \sum_{a_j \in \psi_i} \frac{\left|\left|U_i, U_j\right|\right|_2}{||a_i, a||_2}, \psi_i = \{\tau + 1 \geq \left|\left|a_i, a_j\right|\right|_2 \geq \tau\} \quad (1)$$

$$\tau = \frac{r}{\left|\left|\Lambda(SM_T)\right|\right|_0} \sum_{\tau \in ||\tau,i|| \leq r} \left|\left|\Lambda(SM_{T_\tau})\right|\right|_0 \; ; \; l = 0.5 \min\{width, height\}, \Lambda \rightarrow down\ sampling \quad (2)$$

$$CM_{k,i} \leftarrow \frac{\sum_{\tau=k-1}^{k+1} \sum_{a_{\tau,j} \in \mu\phi} \exp\left(-||c_{k,i}, c_{\tau,j}|| \, 1/\mu\right) \cdot CM_{\tau,j}}{\sum_{\tau=k-1}^{k+1} \sum_{a_{\tau,j} \in \mu\phi} \exp\left(-||c_{k,i}, c_{\tau,j}|| \, 1/\mu\right)} \quad (3)$$

$$\theta = \frac{1}{m \times n} \sum_{k=1}^{n} \sum_{i=1}^{m} || \frac{1}{m} \sum_{i=1}^{m} F(SM_{T_{k,i}}), F(SM_{T_{k,i}})||_1; \; m, n = frame\ numbers \quad (4)$$

$$F\left(SM_{T_i}\right) = \begin{cases} a_i, & SM_{T_i} \leq \epsilon \times \frac{1}{m} \sum_{i=1}^{m} SM_{T_i}; \; \epsilon = filter\ strenght\ control \\ 0, & otherwise \end{cases} \quad (5)$$

3.3 Enhanced Precision

to utilize matrix M and employ RPCA)to decompose it into sparse S and low-level D components. This decomposition serves the purpose of correcting detections by locating the rough foreground, enhancing the accuracy of super-pixel correspondences. The process involves creating feature subspaces for both color and motion, ensuring alignment of super pixels for the subsequent infusion of saliency mapping. The matrix decomposition itself is achieved through specific equations that update the sparse (S) and low-level (D) components iteratively, incorporating singular value decomposition for precision and efficiency. This meticulous approach allows for a nuanced analysis of the video content, addressing both sparse and low-level features in the matrix to refine and enhance the overall accuracy of the system.

$$\min_{M_{cx},S_{cx},\vartheta,A\odot\vartheta} \left||M_c|\right|_* + \left||D_x|\right|_* + \left||A+\vartheta|\right|_2 + \alpha_1\left||S_c|\right|_1 + \alpha_2\left||S_x|\right|; \ ||\cdot||_* \qquad (6)$$

$$\rightarrow nuclear\ norm, A\ is\ position\ matrix$$

$$s.t\ M_c = D_c + S_c,\ M_s = D_s + S_x,\ M_c = SB \odot \vartheta, M_x = SF \odot \vartheta,$$

$$\vartheta = \{E_1, E_2, \ldots., E_n\}, E_i \in \{0,1\}^{m\times m}, E_i 1^K = 1.$$

The variables D_c, D_x are indicative of the color and saliency mapping, respectively, while ϑ represents the permutation matrix. Concurrently, S_x, S_c symbolize the color feature sparse component space and saliency feature space, respectively. This comprehensive set of equations plays a crucial role in the correction of super-pixel correspondences. The iterative interplay of these variables and matrices, guided by the permutation matrix ϑ, facilitates the refinement and alignment of color and saliency mappings. This systematic approach ensures the accurate correction of super-pixel correspondences, contributing to the overall precision and reliability of the proposed system.

4 Result and Discussion

The proposed solution underwent rigorous evaluation and comparison with several state-of-the-art saliency algorithms, including as a baseline reference, OBDL algorithm,'s AWS-D algorithm, the OMCNN-2CLSTM algorithm from, ACL algorithm, SAVC algorithm by , and's Xu et al. The chosen database for experimentation aligns with the base paper, utilizing a high-definition eye-tracking database accessible on GitHub at https://github.com/spzhubuaa/Video-based-Eye-Tracking-Dataset. The dataset comprised 10 video sequences, each with three different resolutions (1920×1080, 1280×720, and 832×480). The performance evaluation encompassed five global metrics, namely AUC, SIM, CC, NSS, and KL. The XU algorithm, resembling HEVC, demonstrated superior saliency detection in less complex images but faced challenges with intricate inputs. In contrast, our proposed algorithm exhibited outstanding performance, surpassing other algorithms in terms of compression efficiency and precision.

4.1 Result analysis

The fixation maps, as depicted within Table 1, provide a visual representation of the saliency results. Subsequently, These visual representations offer a comparative insight into the effectiveness of each algorithm in capturing salient regions. Figure 1 further illustrates the comprehensive evaluation and comparison of saliency algorithms. The graph highlights the performance metrics across various algorithms, providing a visual overview of their comparative strengths and weaknesses. The proposed framework consistently outperformed its counterparts, establishing its efficacy in achieving robust saliency detection across diverse video sequences and resolutions.

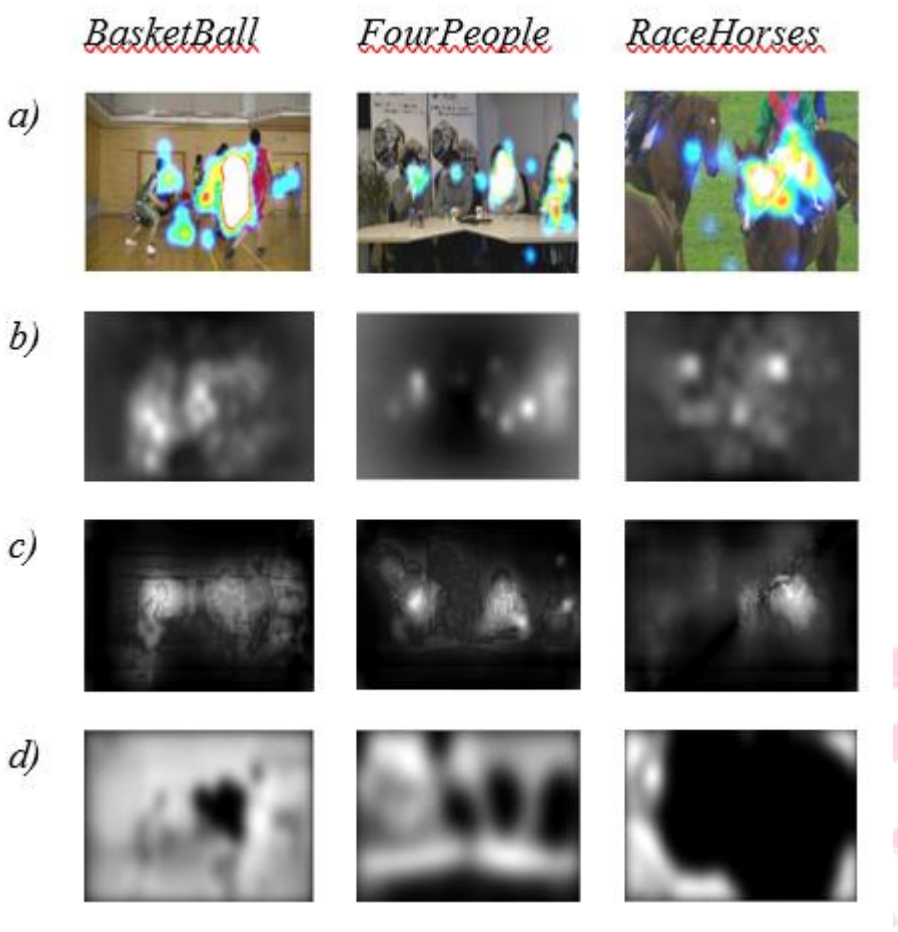

Table 1: The following results for saliency algorithms used; a)Fixation Maps; b) XU[40]; c)Base Paper[34]; d)Proposed Algorithm

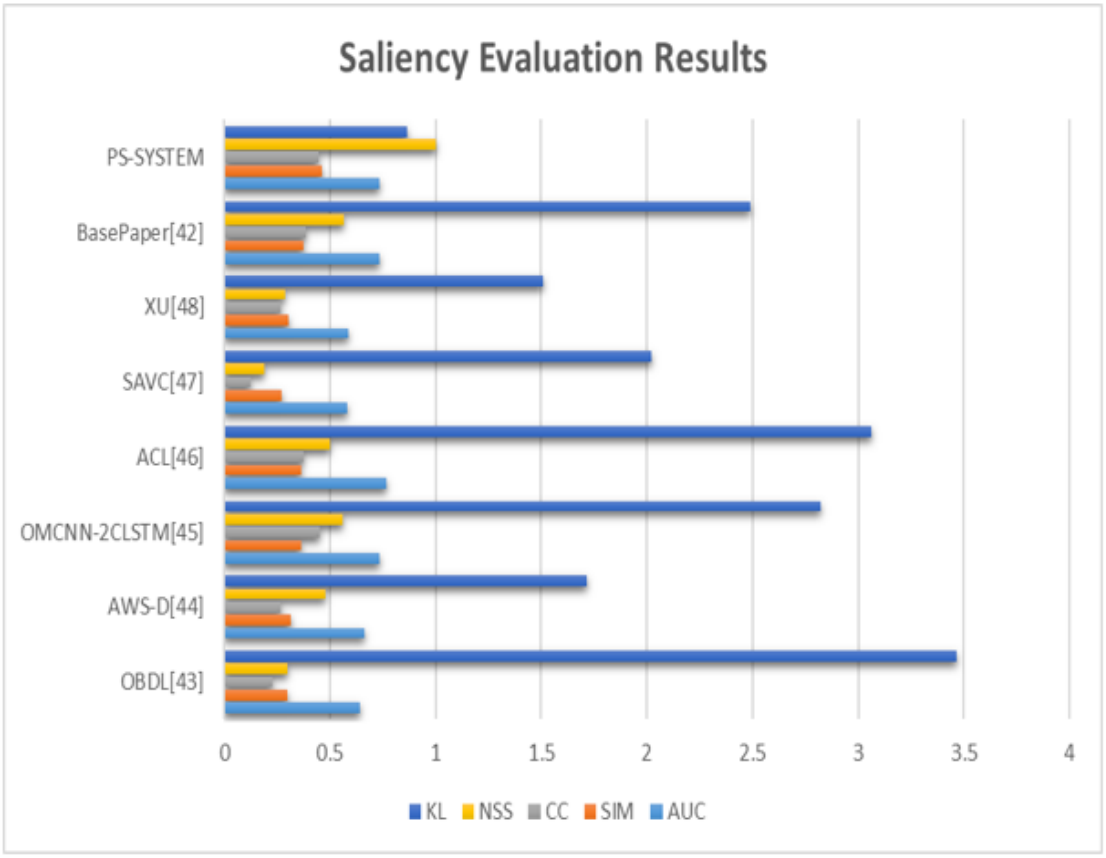

Fig 2 comparison graph

4.2 Comparative analysis

The experimental comparison of saliency algorithms, including XU, OBDL, AWS-D, OMCNN-2CLSTM, ACL, and SAVC, highlights distinct strengths and weaknesses. XU excels in less complex images but struggles with intricate scenes. Other algorithms exhibit varying precision and sensitivity. Remarkably, the proposed framework outperforms its counterparts, demonstrating superior compression efficiency and precision across diverse video sequences and resolutions. The comparative graph visually underscores the consistent excellence of the proposed framework in saliency detection accuracy.

5 Conclusion

The proposed research RSTA shows a dynamic saliency modeling system, integrating long-term inter-batch information and color contrast computation, demonstrates enhanced robustness. Leveraging background and foreground appearance models, along with meticulous handling of RGB history, ensures accurate saliency adjustment. Spatio-temporal gradient mapping and matrix decomposition through RPCA contribute to precise contrast-based saliency mapping. Experimental comparisons highlight the system's superior performance across diverse video sequences and resolutions, surpassing state-of-the-art algorithms in compression efficiency and precision. The consistent excellence of the proposed algorithm affirms its efficacy in achieving robust and accurate saliency detection, showcasing its potential impact on computer vision and video analysis applications. Future endeavors may explore further optimizations to elevate the system's capabilities.

6 Acknowledgement

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.

References

1. Achanta R, Hemami S, Estrada F, Susstrunk S. Frequency-tuned saliency detection model. CVPR: Proc IEEE; 2009. p. 1597–604.
2. Cerf M, Paxon Frady E, Koch C. Faces and text attract gaze independent of the task: Experimental data and computer model. J Vis. 2009;9(12):1–15.
3. Cerf M, Harel J, Einha¨user W, Koch C. Predicting human gaze using low-level saliency combined with face detection. In Platt JC, Koller D, Singer Y, Roweis S, editors. Adv Neural Inf Process Syst 2007;20.
4. Li L-J, Fei-Fei L. What, where and who? Classifying event by scene and object recognition. IEEE Int Conf Comput Vis (ICCV); 2007.
5. Scassellati B. Theory of mind for a humanoid robot. Autonom Robots. 2002;12(1):13–24.
6. Marat S, Ho Phuoc T. Spatio-temporal saliency model to predict eye movements in video free viewing. 16th European Signal Processing Conference EUSIPCO-2008, Lausanne: Suisse; 2008.
7. Ma Y, Zhang H. A model of motion attention for video skimming. Proceedings of IEEE, ICIP, Vol. 1, pp. 22–25; 2002.
8. Shan L, Lee MC. Fast visual tracking using motion saliency in video. Proceedings of IEEE, ICASSP. Vol. 1, pp. 1073–1076; 2007.
9. Peters RJ, Itti L. Beyond bottom-up: incorporating task-dependent influences into a computational model of spatial attention. In: Proceedings of IEEE, CVPR; 2007, p. 1–8.
10. Schu¨tz AC, Braun DI, Gegenfurtner KR. Object recognition during foveating eye movement. Vis Res. 2009;49:2241–53.
11. F. Zhou, S. Kang, and F. Michael, “Time-mapping using space-time saliency,” in IEEE Conference on Computer Vision and Pattern recognition, 2014, pp. 3358–3365.
12. Z. Liu, X. Zhang, S. Luo, and O. L. Meur, “Superpixel-based spatiotemporal saliency detection,” IEEE Transactions on Circuits and Systems or Video Technology, vol. 24, no. 9, pp. 1522–1540, 2014.
13. Y. Fang, Z. Wang, W. Lin, and Z. Fang, “Video saliency incorporating spatiotemporal cues and uncertainty weighting,” IEEE Transactions on Image Processing, vol. 23, no. 9, pp. 3910–3921, 2014.