Comparison of attention metrics for RGB and event data

Lika AMBROSISHVILI Gnouyadou Romaric MAZNA Dalia HAREB Jean MARTINET
Université Côte d'Azur, I3S/CNRS, 2000 rte des Lucioles, 06900 Sophia-Antipolis, FRANCE

Résumé – Cet article présente une étude de plusieurs mécanismes d'attention dans les données RGB et événementielles, en mettant l'accent sur les métriques d'évaluation de la performance. Nos résultats montrent que les résultats peuvent varier considérablement en fonction de la métrique d'évaluation. Nous notons également que, bien que l'attention cognitive (déscendante) reste relativement stable à travers les modalités, l'attention visuelle (ascendante) est fortement influencée par des indices perceptuels propres à chaque format. Ces résultats mettent en évidence la nécessité d'une exploration rigoureuse dans le choix des métriques d'évaluation lors de la conception des mécanismes d'attention, dépendamment des applications.

Abstract – This paper presents a study of several attention mechanisms in both RGB and event data, with a focus on the performance of evaluation metrics. Our results reveal that the performance can vary significantly depending on the evaluation metric. We also note that while cognitive attention (Top-Down) remains relatively stable across modalities, visual attention (Bottom-Up) is strongly influenced by perceptual cues unique to each format. These insights highlight the need for rigorous exploration in selecting evaluation metrics when designing attention mechanisms, depending on specific applications.

1 Introduction

Attention in visual scenes refers to the ability of a system (biological or artificial) to selectively focus on the most relevant parts of a scene while ignoring less important details. This process helps reduce information overload and improve efficiency. Embedded computer vision systems that need to efficiently process large amounts of visual data (whether RGB or event) while maintaining high accuracy and reliability can benefit from attentional mechanisms to help focus on the most relevant parts of the input data, thereby reducing the computational load and enhancing performance.

Attention mixes bottom-up visual attention driven by saliency and top-down cognitive attention driven by the demands of the current task at hand.

The recent concept of Transformer attention provides a computational mechanism to *attend* selected portions of the input. Yet, the link between human attention and Transformer attention remains unclear. Moreover, while a few computational models for saliency prediction (mostly for RGB data) exist, relatively little research, to our knowledge, has investigated cognitive attention using event data despite its promising potential. We study and compare several human and computational attention mechanisms in RGB and event data. Through a user gaze data collection where participants are shown RGB and event images (pseudo-frames formed with $50ms$ events accumulation) as visual inputs, we collect gaze data in top-down and bottom-up settings. Human visual attention was measured without specific instructions, while cognitive attention was assessed with task-related directives. The resulting heatmaps were compared with existing computational models to examine the alignment between human and computational attention mechanisms.

In this paper, we take a metric-driven approach to address this gap. Our central hypothesis is that evaluating attention mechanisms—whether human or computational—requires a multifaceted view. No single metric can capture all aspects of attention behavior. Therefore, we compare attention maps using five complementary metrics: Structural Similarity Index (SSIM), Intersection over Union (IoU), Pearson Correlation Coefficient (CC), Similarity Index, and Kullback-Leibler Divergence (KLD), each highlighting specific aspects of the comparison.

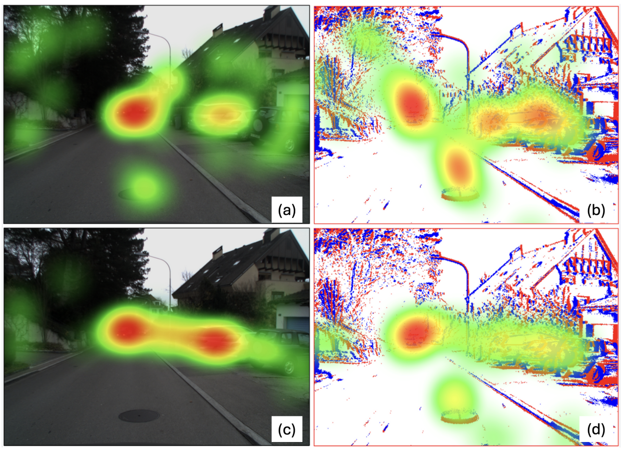

Figure 1 – Illustration of human attention. (a) Visual attention in RGB data, (b) visual attention in event data (c) cognitive attention in RGB data (d) cognitive attention in event data.

The paper aims to illustrate the potential for developing specialized attention models for event cameras in the future, bridging the gap between neuromorphic sensing and human-like visual processing systems.

Fig.1 illustrates the Tobii software experiment output for one of the scenes, showing both RGB and event frames for visual and cognitive attention. These outputs (modified) are later compared to computational models.

2 Comparison metrics

For a comprehensive analysis, we have used 5 different metrics, including location-based and distribution-based ones [2].

Contrary to [2] where saliency maps are in color, our saliency maps are binary and the location-based metrics differ.

2.1 Location-based metrics

The *Structural Similarity Index (SSIM)* [10] compares heatmaps based on structural information by components: luminance (average brightness similarity), contrast (variation similarity), and structure (texture and pattern alignment between heatmaps). SSIM is calculated with Eq. 1, where X and Y represent the input heatmaps to be compared. $\bar{X}$ and $\bar{Y}$ are the mean intensity values of the heatmaps X and Y, respectively. σ_X^2 and σ_X^2 denote the variance of intensity values for X and Y, while σ_{XY} is the covariance between the heatmaps. C_1 and C_2 are small constant values that prevent division by zero when the denominator is null.

$$\text{SSIM}(X, Y) = \frac{(2\bar{X}\bar{Y} + C_1)(2\sigma_{XY} + C_2)}{(\bar{X}^2 + \bar{Y}^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)} \quad (1)$$

The *Intersection over Union (IoU)* (or Jaccard index) measures the overlap between two sets or regions (Eq. 2). The value ranges from 0 to 1.

$$\text{IoU}(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|I|}{|U|} \quad (2)$$

2.2 Distribution-based metrics

The *Pearson Correlation Coefficient (CC)* measures the linear relationship between the heatmaps (Eq. 3). It reflects how well the intensity values in one heatmap correspond to the values in another heatmap.

$$CC(X, Y) = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2}\sqrt{\sum_i (Y_i - \bar{Y})^2}} \quad (3)$$

Here, X_i and Y_i represent the intensity values at the i-th pixel in the two heatmaps. A value of $CC(X, Y) = 1$ indicates perfect positive correlation, -1 is a perfect negative correlation, and 0 means no correlation.

The *similarity (SIM)* metric [8] measures the degree of overlap between two heatmaps X and Y by summing the minimum value at each corresponding pixel position (Eq. 4). Both maps are normalized such that the sum of all pixel values equals 1 to ensure that SIM ranges between 0 and 1 (1 indicates a perfect match and 0 means no similarity between heatmaps).

$$\text{SIM}(X, Y) = \sum_i \min(X_i, Y_i) \quad (4)$$

The *Kullback-Leibler Distance (KLD)* [7] [9] quantifies the dissimilarity between two probabilities distributions X and Y by averaging their Kullback-Leibler divergences D_{KL} relative to their mean distribution, as in Eq. 5.

$$KLD(X, Y) = \frac{1}{2}\left[D_{KL}\left(X || \frac{X+Y}{2}\right) + D_{KL}\left(Y || \frac{X+Y}{2}\right)\right] \quad (5)$$

Note that we define and use a new metric, KL Similarity $KLS(X, Y) = 1 - KLD(X, Y)$, instead of the standard KL divergence. This formulation ensures consistency across all metrics, aligning them such that higher values indicate greater similarity. Complementary to the 5 comparison metrics, we calculated a combined score that represents the average of min-max normalized [1] metric values to summarize the data trend.

3 Attention maps

The comparison between the several attention metrics has required the collection of human gaze data with an eye tracker and to adapt of computational models. The study has been carried out using the Stereo Event Camera Dataset for Driving Scenarios (DSEC) [5]. DSEC includes synchronized recordings from two standard RGB cameras and two event cameras. In this work, we do not utilize the stereo information. This dataset enables direct comparison between RGB and event data.

3.1 Human gaze data collection

Data collection was performed using a Tobii Pro Nano eye tracker in conjunction with Tobii Pro Lab software [1].The software was used to present stimuli, run the experiment, and generate fixation heatmaps.

Participants: A total of 24 volunteers participated: 16 males (66.7%) and 8 females (33.3%) aged between 18 and 40 years old (average 24 years, std 5.05 years). Most participants – 19 out of 24 (79%) – are glass wearers. None of the participants had prior experience with eye tracking or vision-related tasks and had never been exposed to event data. Additionally, they were asked about their driving experience: 17 participants (70.8%) had a valid driving license, and 7 participants (29.2%) had not. This information is important since drivers may visually explore a driving scene differently from non-drivers.
Stimuli distribution: The gaze data collection sessions consisted of two parts in sequence: Task 1 for visual attention and Task 2 for cognitive attention (see details below). We used 20 pairs of images from 20 scenes, which amount to a total of 40 stimuli: 20 RGB images and 20 event images (pseudo-frames made of $50ms$ events accumulation). The 20 scenes were split into four parts 1, 2, 3, and 4, which makes, considering the RGB and event modalities, the following eight subsets each containing five images: R1, R2, R3, and R4 for RGB images, and E1, E2, E3, and E4 for event images. The eight subsets have been carefully distributed among participants and tasks for Task 1 and Task 2, as we wanted each participant to be exposed to only half of the stimuli to ensure that no individual saw the same scene in both RGB and event modalities, avoiding potential bias in the gaze such as looking at already familiar objects on a stimuli. The 24 participants were divided into four groups: A, B, C, and D, and performed tasks as follows:

— A: Task 1 on R1 ∪ E1, then Task 2 on R2 ∪ E2
— B: Task 1 on R2 ∪ E2, then Task 2 on R1 ∪ E1
— C: Task 1 on R3 ∪ E3, then Task 2 on R4 ∪ E4
— D: Task 1 on R4 ∪ E4, then Task 2 on R3 ∪ E3

[1] Software version 24.21.435(x64), OS Windows 10.

User experiment setup: Each session began with a calibration process to ensure that the hardware was adjusted to each participant, as variations in height and seating posture affect the gaze angle. Each participant was sitting on average 70cm distance from a 1980x1080 screen. Calibration included looking at 5 exploding dots (in the center and corners), which is a typical Tobii Task Manager calibration process. After the calibration process, each participant was shown a shuffled stimulus set consisting of 5 RGB images and 5 event-framed images. As described earlier, visual attention is unconscious and stimuli-driven. In the first part dedicated to visual attention, participants were asked to freely observe each photo for 7 seconds since studies in the literature [4] show that 5-7 seconds is typically enough to capture initial visual attention patterns. In the second half, subjects were exposed to cognitive attention tasks. Since top-down cognitive attention is task-driven, participants were given the task to count out loud four-wheeled vehicles (cars, trucks, and buses) in the scene. The experiment observer wrote down the numbers, pretending that the counted numbers were the key element of the experiment, which encouraged participants to maintain a consistent cognitive effort across all trials. Similarly to the previous task, each stimulus was presented for 7 seconds and consisted of 5 RGB images and 5 event images. Each participant spent around 3 to 5 minutes in total on the experiment, depending on the calibration time and reading speed for instructions.

3.2 Saliency maps from Itti-Koch model

Saliency maps were generated with the Python library PySaliencyMap[2]. The Itti-Koch model extracts intensity, color, and orientation features using Gaussian pyramids and Gabor filters. Then, it computes center-surround contrasts to highlight differences across spatial scales, followed by normalization. The final salience maps are obtained by combining these processed features.

3.3 Transformer feature maps

To compare human cognitive attention with Transformer-based attention, we used the DETR model [3]. Specifically, we extracted output feature maps from its Transformer architecture, which consists of an encoder with self-attention and a decoder incorporating both self-attention and cross-attention. The model, trained on ImageNet, was used to detect vehicles in RGB frames from the DSEC dataset.

For this purpose, we use a Transformer implementation by Facebook Research Group[3] that generates predictions and allows to visualize the attention of the model.

However, Transformers trained on event-based data lack the original decoder component of the Transformer architecture. When using RVT model [6] in our experiments – aiming to extract the same Transformer feature maps as in DETR – we encountered some challenges. In DETR, the feature maps resemble attention maps, highlighting detected objects. RVT diverges by employing a multi-stage encoder that integrates Transformers with both local and dilated global self-attention for spatial feature interaction, eliminating the need for a decoder altogether. To address this challenge, we opted to consider the model's output directly to generate the attention maps.We created a binary mask by initializing an image with all pixels set to 0 and assigning a value of 1 to pixels within the predicted bounding boxes. Fig.2 illustrates a visual representation of attention maps across different modalities for a single scene.

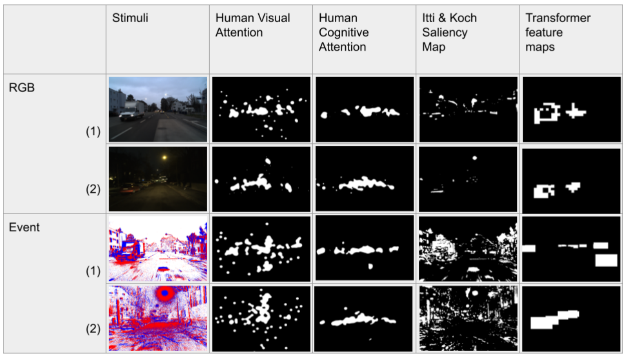

Figure 2 – Sample input data (first column) and their corresponding binarized heat maps collected from participants' gaze tracks (second and third columns) and generated with saliency and Transformers models (last two columns).

4 Results and discussion

The main results of the study are given in Fig. 3 and Tab. 1 with the detailed metrics values.

While SSIM yields high performance, IoU on the other side shows relatively poor performance. SSIM good performance could be explained by its sensitivity to local structure and luminance patterns. It indicates that the local structures between compared heatmaps are quite similar.

The poor IoU performance suggests that even though the local structures might be similar, the predictions and ground-truths do not overlap well, especially in binary maps where small differences can affect significantly the IoU score.

	SSIM	CC	IoU	SIM	KLS
Human attention in RGB: TD vs BU	0.7628	0.5048	0.1600	0.0321	0.9968
Human attention in Event: TD vs BU	0.7143	0.4238	0.1305	0.0301	0.9934
Human BU attention: RGB vs event	0.6808	0.4318	0.1401	0.0402	0.9978
Human TD attention: RGB vs event	0.8261	0.6104	0.2064	0.0306	0.9996
RGB: Human BU vs Itti-Koch saliency	0.9040	0.3960	0.2559	0.0152	0.9955
Event: Human BU vs Itti-Koch saliency	0.8724	0.3079	0.1936	0.0352	0.9744
RGB: Human TD vs DETR attention	0.8502	0.3115	0.2151	0.0170	0.9971
Event: Human TD vs RVT output	0.9398	0.4997	0.3487	0.0224	0.9916

Table 1 – Averaged performance of metrics.

As KLD quantifies the difference between two probability distributions, our KLS high score indicates that the distribution of saliency values is similar across the predictions and the ground truths.

The very low SIM score means that there is near to no agreement when the distributions are compared directly as described earlier (Eq. 4). This reinforces the idea that the distributions of predicted salient areas do not closely match the ground truths, despite possible structural similarities.

[2] https://github.com/akisatok/pySaliencyMap, accessed in February 2025.

[3] URL: https://colab.research.google.com/github/facebookresearch/detr/blob/colab/notebooks/detr_attention.ipynb, accessed in February 2025.

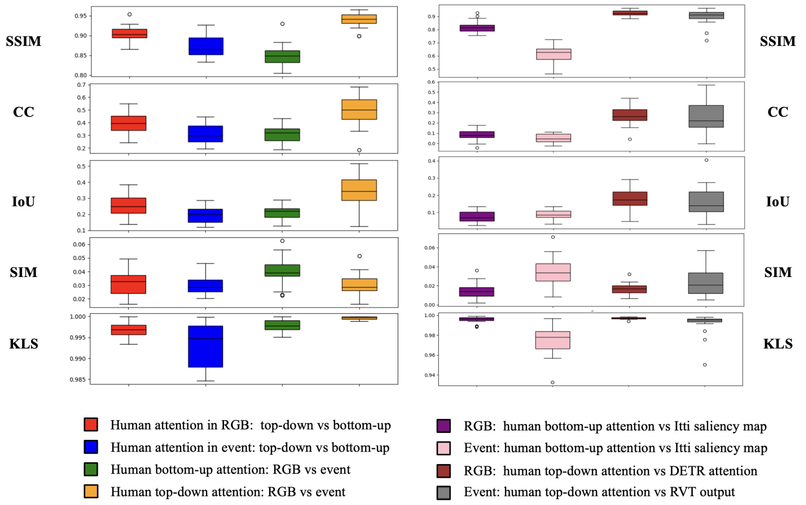

Figure 3 – Heatmaps comparison with 5 different metrics: SSIM, CC, IoU, SIM, KLS.

The CC metric yields moderate results, potentially capturing a level of similarity that lies between exact spatial overlap (as measured by IoU) and broader distributional differences (as measured by KLD and SIM). This suggests that while certain aspects of the overall distribution are reasonably aligned, notable discrepancies remain in critical regions.

Overall the high SSIM indicates that on a structural level, the predicted saliency maps captured some of the essential patterns and edges. However, the low IoU tells us that when it comes to the exact placement or extent of the salient regions, the prediction is lacking.

The disparity observed in distribution-based metrics further confirms that the overall distribution of salient pixels even if similar (high KLS), is not matching well with the ground truth (low SIM).

5 Conclusion

This study provides an analysis of the performance of evaluation metrics when comparing attention mechanisms in both RGB and event data, specifically human visual attention, human cognitive attention, and computational models. The findings highlight several differences in performance depending on the evaluation metric.

This study suggests that the choice of metrics should be examined carefully depending on the applications. For applications where the exact location and extent of the salient region is crucial, a low IoU and distribution discrepancies (KLS, SIM) are concerning. However, if the task can tolerate some spatial misalignment as long as the local structure is maintained, a high SSIM might be a positive sign. This work lays the foundation for future research in neuromorphic vision, where event cameras and bio-inspired attention models could enhance real-time processing in robotics, autonomous vehicles, and surveillance applications.

Acknowledgements: This work was supported by the project NAMED (ANR-23-CE45-0025-01) of the French National Research Agency (ANR).The authors would like to thank Teresa Colombi for the constructive discussions.

References

[1] Peshawa Jamal Muhammad Ali. Investigating the impact of min-max data normalization on the regression performance of k-nearest neighbor with different similarity measurements. *ARO-The Scientific Journal of Koya University*, 10(1):77–83, 2022.

[2] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What Do Different Evaluation Metrics Tell Us About Saliency Models? *IEEE TPAMI*, 41(3):740–757.

[3] Nicolas Carion, Francisco Massa, et al. End-to-end object detection with transformers. In *ECCV*. Springer, 2020.

[4] S. Egner, S. Reimann, R. Hoeger, and W. H. Zangemeister. Attention and information acquisition: Comparison of mouse-click with eye-movement attention tracking. *Journal of Eye Movement Research*, 11(6):10.16910/jemr.11.6.4, 2018.

[5] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE RAL*, 6(3), 2021.

[6] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. In *ICML*, 2015.

[7] David Pinto, José-Miguel Benedí, and Paolo Rosso. Clustering narrow-domain short texts by using the kullback-leibler distance. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 611–622. Springer, 2007.

[8] Michael J Swain and D H Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[9] Flemming Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46(4):1602–1609, 2000.

[10] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.