Trends, Applications, and Challenges in Human Attention Modelling

G. Cartella[1], M. Cornia[1], V. Cuculo[1],
A. D'Amelio[2], D. Zanca[3], G. Boccignone[2], R. Cucchiara[1]

[1] *University of Modena and Reggio Emilia,* [2] *University of Milan,*
[3] *Friedrich-Alexander-Universitat Erlangen-Nurnberg*

ITSERR
Associated project to RESILIENCE

AImage Lab

IJCAI JEJU 2024

1. Motivation of the Survey

Integration of Human Attention in AI Models. The perspective of the interplay between deep learning-based applications and visual attention is lacking in most recent reviews of the field. Hence, we offer insights into the integration of human attention into deep-learning models to tackle challenges related to images/videos, text, or multimodal data.

New Research Directions. Our survey identifies the key challenges inspiring future research directions to solve data scarcity and privacy issues.

2. Introduction: Human Attention Modelling

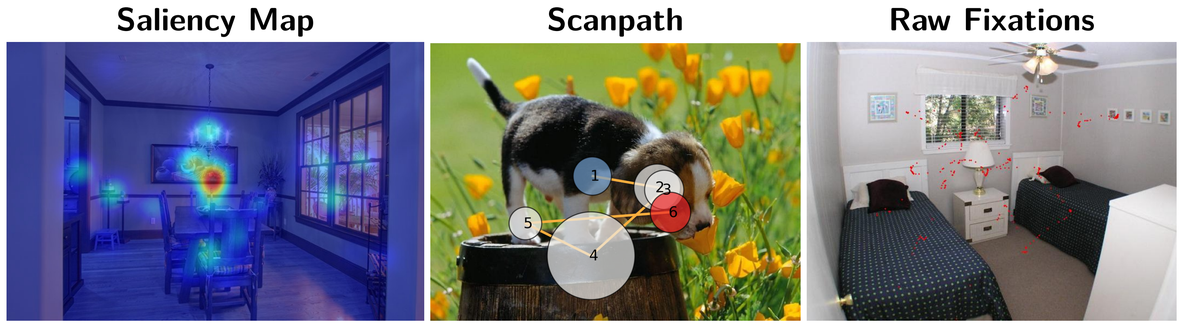
Visual attention enables humans to rapidly analyze complex scenes and devote their limited cognitive resources to the most attractive regions.

Two types of attention:

- **Bottom-up:** Prioritization is guided by the low-level processing of the visual stimulus.
- **Top-down:** It refers to the voluntary allocation of attention to certain features, objects or regions and is usually task-guided (*e.g.* visual search).

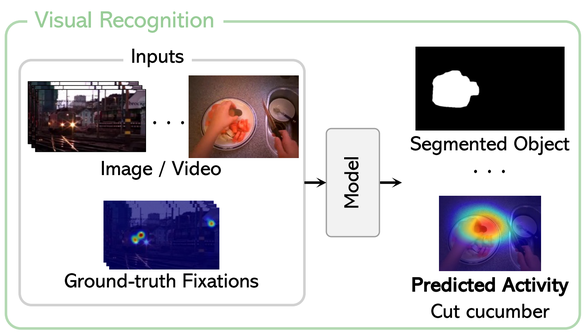
Human Attention has been exploited within different modalities including **Image**, **Video**, **Audio** and **Text**. It can be represented in different forms:

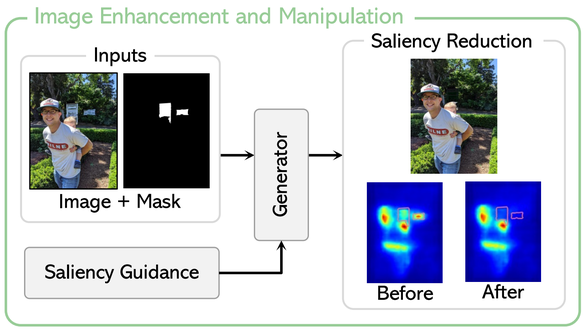
Saliency Map | **Scanpath** | **Raw Fixations**



3. Image and Video Processing

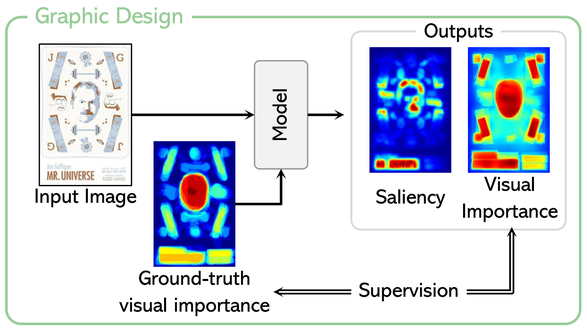
Human gaze data assists AI models in **detecting** salient objects in videos, **segmenting** videos without supervision, and **recognizing** activities in egocentric video sequences.



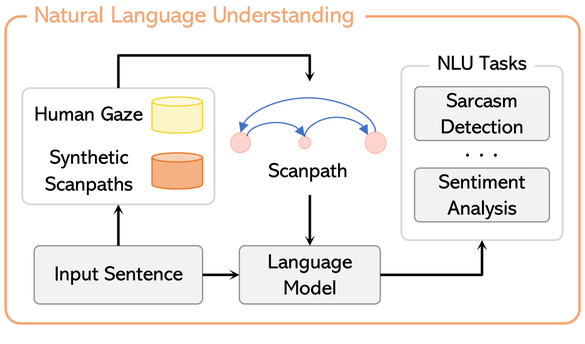


Human saliency is adopted as guidance to manipulate and **enhance** images. The aim is to automatically reduce visual distraction while increasing the attractiveness of the desired objects within the scene.

Visual attention allocation in graphic designs can be interpreted as a proxy for the perceived relative **importance** of design elements.



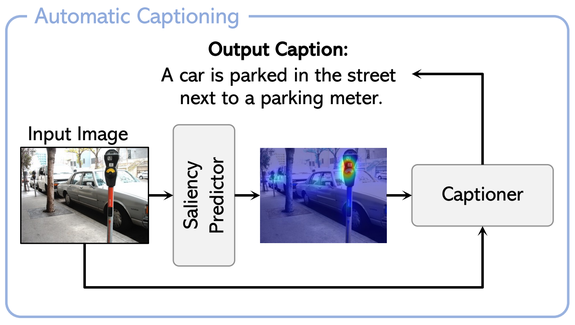
4. Language Modelling



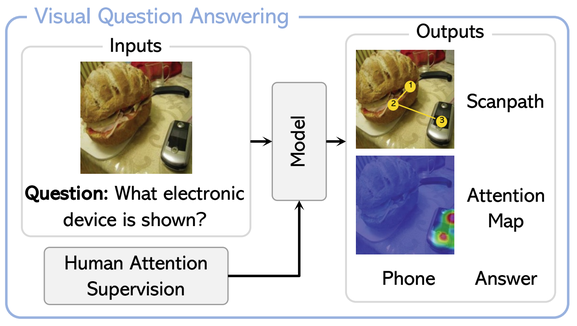
Machine Reading Comprehension: Eye movement patterns are used to infer the reader's native language.

Natural Language Understanding: Models generate scanpaths from existing eye-tracking corpora, improving performance on tasks such as sentiment analysis and sarcasm detection.

5. Vision-and-Language Applications

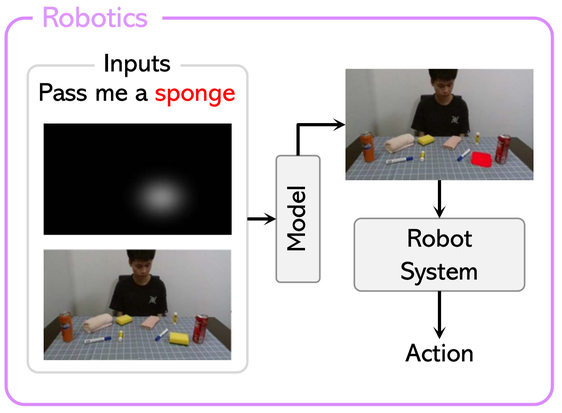


Raw fixations or saliency maps can be combined with the language model to guide the **caption** generation.



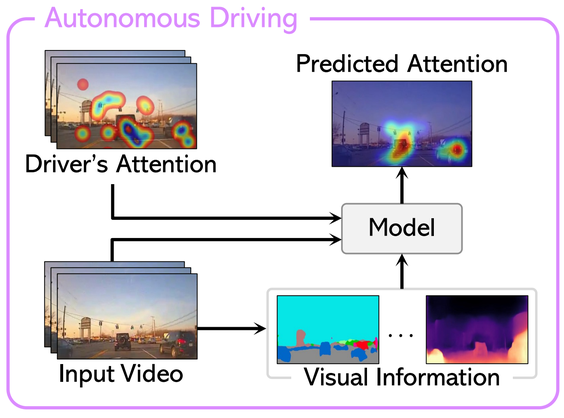
Human gaze guides VQA models to **generate** human-like attention maps and promoting **reasoning** behaviour.

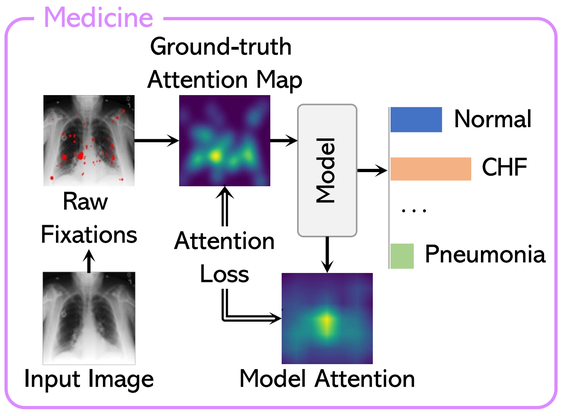
6. Domain-Specific Applications



In collaborative robotics, robot assistance can be enabled by individuals directing their gaze towards a specific target they wish to grasp or **manipulate**. **Control systems** able to accommodate eye-gaze input modalities have been developed.

Inverse reinforcement learning can be adopted to learn an attention policy treating each fixation point as a potential source of **reward**. Some works modelled the human peripheral vision and introduced a gaze detection model augmented by the **scene semantics**.





Human gaze emerges as a natural way to capture visual attention during the **diagnosis process**. Eye movements from an expert are recorded and raw fixations are converted into a visual attention map to guide model attention towards the most discriminative regions.

7. Current Challenges and Future Research Directions

⚠️ **Data Scarcity:** Collecting human gaze data is expensive.

🚀 **Wearable Devices:** The use of wearable devices such as smart glasses and augmented reality headsets would allow large amounts of data to be collected efficiently and in an ecological setting.

⚠️ **Privacy Issues:** Data acquisition through wearable devices raises a new challenge concerning ethical and privacy-aware collection and sharing.

🚀 **Synthetic Data:** In the NLP community, the integration of synthetic scanpath generation with a scanpath-augmented language model has shown promising results, eliminating the need for human gaze data. However, minimal exploration has been undertaken in computer vision.

[1] S. Deng et al., *Pre-Trained Language Models Augmented with Synthetic Scanpaths for Natural Language Understanding*, in EMNLP, 2023.
[2] S. Miangoleh et al., *Realistic Saliency Guided Image Enhancement*, in CVPR, 2023.
[3] C. Fosco et al., *Predicting visual importance across graphic design types*, in ACM UIST, 2020.
[4] M. Ilaslan et al., *GazeVQA: A Video Question Answering Dataset for Multiview Eye-Gaze Task-Oriented Collaborations*, in EMNLP, 2023.
[5] B. Wang et al., *GazeGNN: A Gaze-Guided Graph Neural Network for Chest X-Ray Classification*, in WACV, 2024.

Have a look at our awesome repo!