Benchmarking human visual search computational models in natural scenes: models comparison and reference datasets

Fermín Travi[12*], Gonzalo Ruarte[12*], Gastón Bujía[12], Juan E. Kamienkowski[13]

*both authors contributed equally to the present work

1. Laboratory of Applied Artificial Intelligence, ICC, FCEyN, UBA-CONICET
2. Department of Computing (DC), FCEyN, UBA
3. Master in Data Mining & Knowledge Discovery, FCEyN-UBA

FerminT/VisualSearchBenchmark fermintravi@gmail.com

SAN2021 XXXVI ANNUAL MEETING OCTOBER 18-22

Background

Visual search in natural scenes consists of looking for a target object in an image which resembles everyday life.

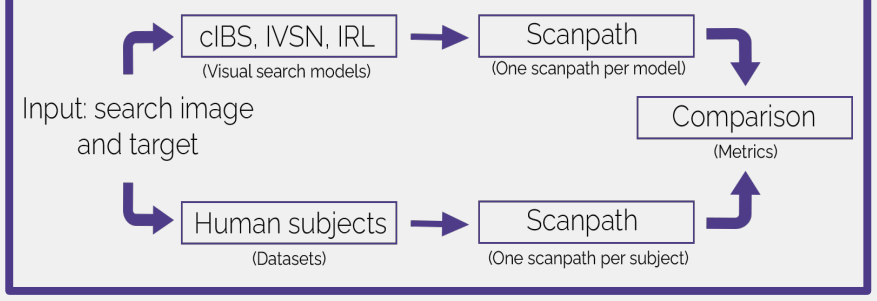
- Humans carry out this task by performing successive eye movements **(saccades)**. The intervals of time between saccades are called **fixations**. This is when information is incorporated. A **scanpath** is an ordered list of fixations.

Example of a scanpath on a search image

- Nowadays, several algorithms are able to predict gaze positions during simple observation, but few models attempt to simulate human behaviour during visual search in natural scenes.
- Moreover, these models vary widely in their design, and exhibit differences in the datasets and metrics with which they were evaluated.
- Therefore, there is a need for a reference point, on which each model can be evaluated, and from where potential improvements to the algorithms can be derived.

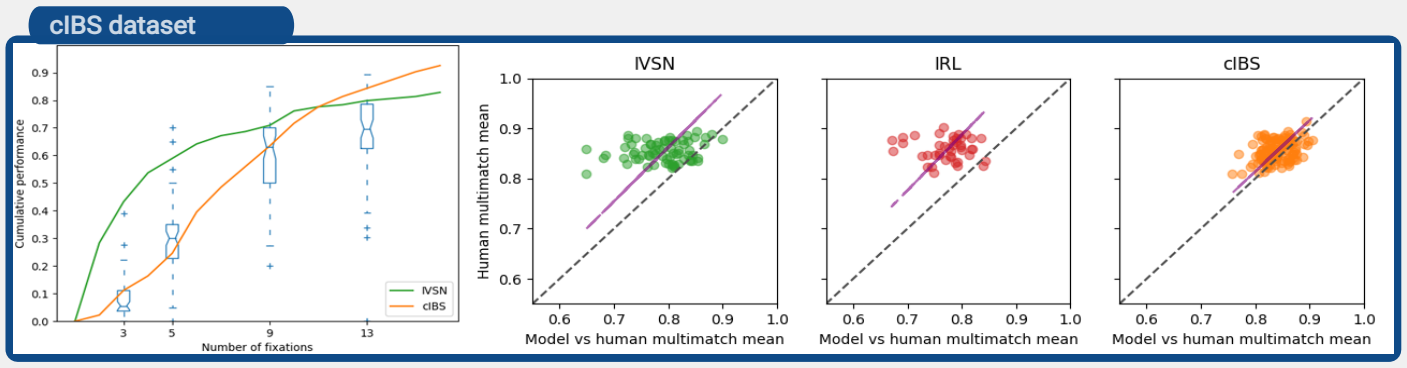
Methods

- **Three** different visual search **models**, each with its own architecture, and **four** publicly available **datasets** are considered.
- Each **dataset** comprises a set of search images and **human subjects' scanpaths** on those images.

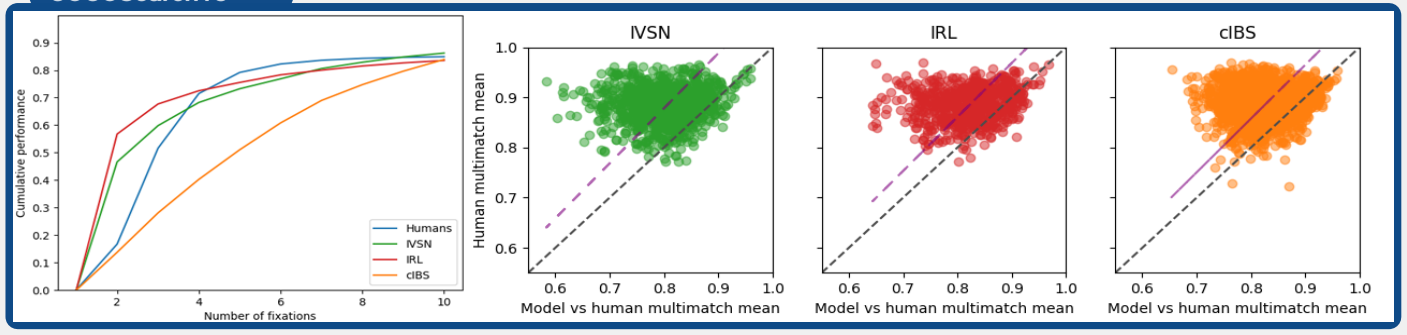


Results

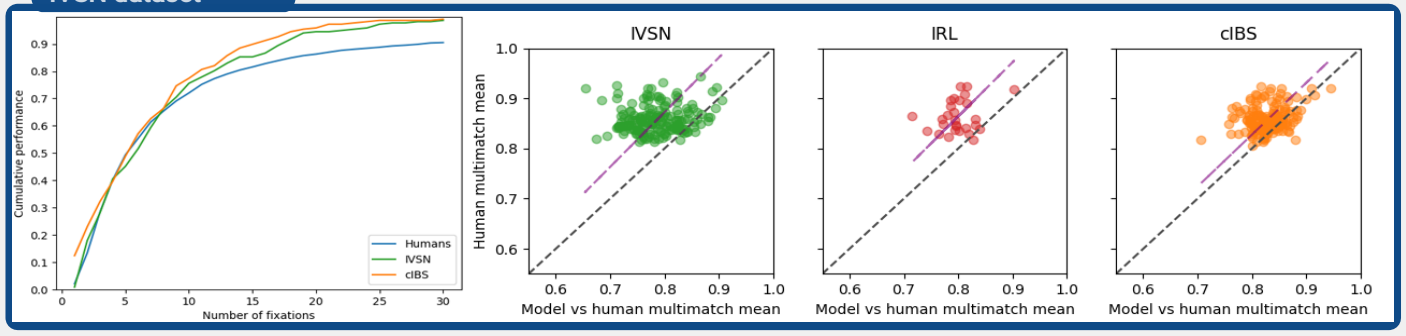
cIBS dataset



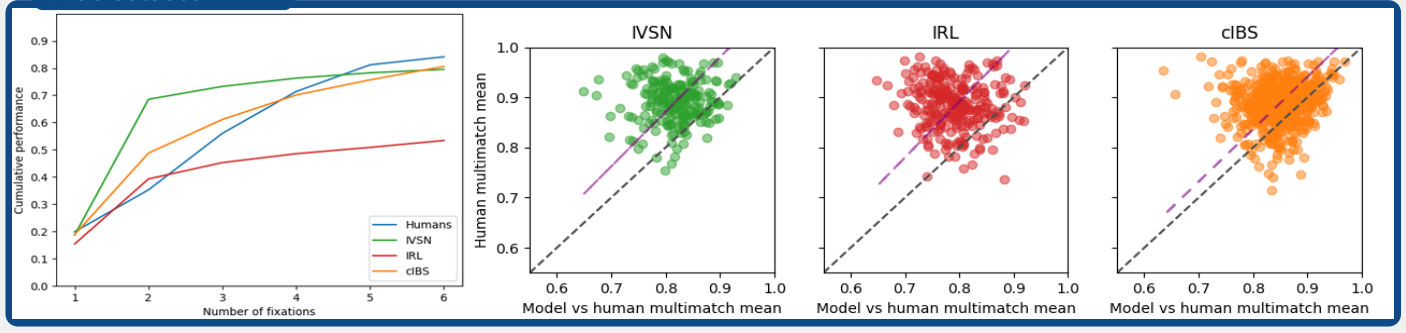
COCOSearch18



IVSN dataset



MCS dataset



Models

cIBS[1]

- **Bayesian model**. Each eye movement is thought of as a decision on where to look next.
- The **prior** is a saliency map (computed with DeepGaze II).
- The **likelihood** is computed through a visibility map and a target similarity map
- The target similarity map is computed via **cross-correlation**. Here, we consider a variation of the model, using **IVSN's attention map**.
- This enables the model to capture **object invariance**, instead of just templates.

IVSN[2]

- **Based on a CNN** (VGG16). It attempts to simulate humans' visual cortex.
- **Zero-shot:** it was not trained on visual search data. Captures **object invariance**.
- It builds an **attention map** by forwarding both the search and target image in the CNN.
- It performs **greedy search** on the attention map. There is no accumulation of information across saccades.
- There is **no model of the fovea** (shows constant acuity across the entire visual field).

IRL[4]

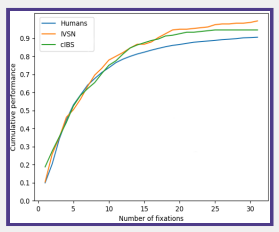
- **Based on Inverse Reinforcement Learning**. It was trained on COCOSearch18.
- Makes use of **GAIL (Generative Adversarial Imitation Learning)**
- Incorporates **context** by preprocessing the search image with Detectron2, computing its **panoptic segmentation.**
- Performs **categorical visual search**. Thus, it is constrained to the 18 object categories it was trained on.
- It could only be tested on a **small subset** of the cIBS and IVSN datasets.

Human subjects

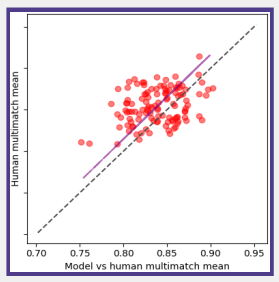
	#Subjects	#Images	#Scanpaths*	Average scanpath length*	People	Exact Target	Distractors	Colour	Image res.
cIBS[1]	57	134	~3.9K	5.5	✗	✓	✓	✗	768x1024
IVSN[2]	15	217	~3.0K	8.16	✓	✗	✗	✗	1024x1280
MCS[3]	27	2203	~2.9K	2.8	✗	✗	✗	✓	508x564
COCO Search18[4]	10	2489	~21K	3.5	✗	✗	✗	✓	1050x1680

Only scanpaths where the target was found are taken into account

Metrics



- **Cumulative performance** computes the ratio of targets found (vertical axis) for a given number of fixations (horizontal axis).
  It is a measure of **efficiency**.
  In the **cIBS dataset**, due to the design of the experiment, subjects' performance are displayed as boxplots.



- **Multimatch**[5] compares two scanpaths in multiple dimensions. In the vertical axis, each human's scanpath is compared against the other humans and then averaged. In the horizontal axis, the model's scanpath is compared against every human and then averaged. Thus, **the closer the points are to the diagonal, the more similar the model is to humans' behaviour**.
  It is a measure of **human similarity**.

Conclusions

- **IRL model failed to generalize** well to other datasets, performing poorly beyond its own dataset.
- **IVSN model** showed **great search efficiency** and was not particularly affected by distractors, but **failed to capture human patterns of behaviour**.
- **cIBS model was the most consistent with human behaviour**, both in efficiency and scanpath similarity. Nonetheless, it still failed to be as efficient as human subjects in the COCOSearch18 dataset.
- The **greatest dissimilarity between human subjects and computational models** occurred in images where **capturing the target object's context** was paramount. Even though the IRL model made a step forward in this direction, its approach is not generalizable. **Future research** must take this factor into account.
- Surprisingly, **human subjects did not follow common distractors** (such as human faces) while looking for an object, but cIBS model did.
- The present work shows an **urgent need for a common set of metrics and data** for the development of more general computational visual search models in natural scenes.

References

[1] Sclar, M., Bujia, G., Vita, S., Solovey, G., Kamienkowski, Juan. (2020). Modeling human visual search: A combined Bayesian searcher and saliency map approach for eye movement guidance in natural scenes. https://arxiv.org/pdf/2009.08373.pdf
[2] Zhang, M., Feng, J., Ma, K.T. et al. Finding any Waldo with zero-shot invariant and efficient visual search. Nat Commun 9, 3730 (2018).
[3] Predicting Goal-directed Attention Control Using Inverse-Reinforcement Learning. G.J. Zelinsky, Y. Chen, S. Ahn, H. Adeli, Z. Yang, L. Huang, D. Samaras and M. Hoai, 2020. arXiv preprint arXiv:2001.11921.
[4] Benchmarking Gaze Prediction for Categorical Visual Search, G.J. Zelinsky, Z. Yang, L. Huang, Y. Chen, S. Ahn, Z. Wei, H. Adeli, D. Samaras and M. Hoai. CVPR Workshops 2019.
[5] Jarodzka, H., Holmqvist, K., & Nyström, M. (2010, March). A vector-based, multidimensional scanpath similarity measure. In Proceedings of the 2010 symposium on eye-tracking research & applications (pp. 211-218).