Búsqueda visual sobre imágenes naturales: incorporando ruido como modelado de la variabilidad humana.

Mateo Feldman (0009-0009-4522-7667, mfeldman@dc.uba.ar)[1,2],

Juan E. Kamienkowski (0000-0002-5725-6539, juank@dc.uba.ar)[1,2,3],

Gonzalo Ruarte (0009-0000-2013-782X, gruarte@dc.uba.ar)[1,2]

[1] *Laboratorio de Inteligencia Artificial Aplicada, Instituto de Ciencias de la Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires - CONICET, Argentina;*
[2] *Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina;*
[3] *Maestría de Explotación de Datos y Descubrimiento del Conocimiento, FCEyN-FI, UBA, Argentina*

Resumen. La búsqueda visual es fundamental en la interacción cotidiana entre los humanos y su entorno que involucra movimientos oculares secuenciales. Diversos modelos computacionales intentan emular este comportamiento imitando los mecanismos cognitivos involucrados en la percepción de imágenes naturales. A pesar de los avances en neurociencia, la búsqueda visual sigue siendo un proceso complejo y no completamente caracterizado. Un aspecto clave de la cognición humana que suele pasarse por alto es el papel del ruido, especialmente en tareas de exploración y toma de decisiones. Distintas fuentes de ruido generan variabilidad en los tiempos de respuesta, provocan errores y podrían explicar diferencias individuales entre participantes. En este trabajo proponemos una versión modificada del modelo Entropy Limit Minimization (ELM) que, si bien es determinista, introduce aleatoriedad mediante semillas derivadas tanto de la imagen de entrada como del participante para capturar mejor dicha variabilidad humana. Exploramos distintas estrategias de inyección de ruido, incluyendo agregado de ruido a la prior de la imagen, al mapa de ganancia de información y al desplazamiento aleatorio de las fijaciones. La evaluación con métricas del benchmark VISIONS mostró que alterar el prior fue la estrategia que más se acercó a la variabilidad humana.

Palabras clave: Búsqueda visual, Ruido, Variabilidad humana

Visual search on natural images: Incorporating noise as a model of human variability.

Abstract. Visual search is crucial in daily human interaction with the environment, involving sequential eye movements. Computational models attempt to replicate this behavior by mimicking cognitive mechanisms used during natural image perception. Despite advances in neuroscience, visual search remains complex and not fully characterized. One key aspect of human cognition often overlooked is the role of noise, particularly in exploration and decision-making. Different sources of noise cause variability in response times, provoke mistakes, and may also explain individual differences between participants. We propose a modified version of the Entropy Limit Minimization (ELM) model to better capture human variability in visual search. Although deterministic, our model introduces randomness through seeds derived from both the input image and the participant. We explore several noise injection strategies to the model, including adding noise to the image prior, the information gain map, and the overshooting of saccades via random offsets. To evaluate variability, we use metrics from the VISIONS benchmark, repeating model runs with different seeds. Notably, adding noise in the prior of the model yielded the closest approximation to human variability. However, the results show that adding different types of noise helps the model approximate human variability in most cases.

Keywords: Visual search, Noise, Human Variability.

1 Introducción

La búsqueda visual es una tarea que consiste en encontrar un objeto particular en un entorno visual. El objeto que se busca se denomina "*target*", mientras que cualquier otro objeto que no sea el buscado se denomina "*distractor*" (Wolfe et al., 2008). Este tipo de búsqueda es fundamental para la vida diaria de los seres humanos. Desde actividades simples como encontrar las llaves en un escritorio desordenado, hasta tareas más críticas como localizar señales de peligro mientras se conduce.

La búsqueda visual está basada sobre los movimientos oculares. Dentro de los movimientos oculares se pueden identificar las fijaciones y las sacadas. Una **fijación** (*fixation*) ocurre cuando el ojo se mantiene enfocado en un lugar específico. Mientras esto sucede, nueva información sobre el entorno visual es incorporada. El movimiento rápido del ojo de una fijación a otra se denomina **sacada** (*saccade*), y no incorpora nueva información. La secuencia ordenada de fijaciones y sacadas se denomina **scanpath**.

Desarrollar modelos computacionales que modelen este proceso puede tener aplicaciones clave en distintas áreas de tecnología y diseño. En inteligencia artificial, por ejemplo, estos modelos pueden mejorar sistemas de visión por computadora y robots, haciéndolos más eficientes en la navegación en entornos desconocidos.

Se han propuesto diversos modelos basados en distintos enfoques como Redes Neuronales Profundas (Zhang et al., 2018), Aprendizaje por Refuerzos

(Zelinsky et al., 2021), y Bayesiano (Bujia et al. 2022). Estos enfoques fueron comparados en el punto de referencia VISIONS (Travi et al., 2022). A partir de los resultados se logró mejorar el enfoque Bayesiano introduciendo redes neuronales en el procesamiento de las imágenes (bottom-up). Recientemente, el proceso de decisión de la siguiente sacada (top-down) fue reemplazado por un modelo Entropy Limit Minimization (ELM) (Najemnik et al., 2009, Bujia et al, 2025, Ruarte et al., 2024).

Este manuscrito tiene como objetivo modelar la variabilidad humana en una tarea de búsqueda visual mediante la introducción de distintos tipos de ruido en distintos componentes del modelo ELM, generando así, una noción de incerteza realista.

2 Metodología

2.1 Conjunto de datos

Se utilizó un subconjunto de 453 imágenes de un conjunto de datos compuesto por 912 imágenes naturales en el exterior (600 x 800px) para la experimentación (Ehinger et al, 2009). Para el presente trabajo se eligió este subconjunto por un límite de recursos computacionales para la experimentación, pero será extendido al total. En todas las imágenes 14 participantes debieron buscar una persona (target). El dataset incluye las imágenes y los scanpaths realizados por cada participante. Cada scanpath incluye la secuencia de tuplas (x,y) de cada fijación y si el participante logró encontrar el objetivo.

2.2 Modelo base: Entropy Limit Minimization (ELM)

Se utilizó el modelo ELM, que consiste en elegir iterativamente la siguiente fijación que maximiza la ganancia de información esperada y minimiza la incertidumbre sobre la ubicación del objetivo. Para ello, calcula un mapa de información esperada considerando la detectabilidad del objetivo en cada posible fijación desde la fijación actual (mapa de visibilidad en la Fig 1.A.), la probabilidad actual de que el objetivo esté en cada ubicación (prior) y la similitud del objetivo (mapa de target similarity) sobre cada posible fijación, eligiendo aquella que maximice la información ganada de encontrar el objetivo.

2.3 Fuentes de ruido

Se analiza la incorporación de ruido en tres componentes distintos del modelo (Fig. 1A): 1. Sobre el prior inicial: cuando el modelo hizo pocas fijaciones, no tiene información sobre la imagen, por lo que añadir ruido al prior implicaría modificar significativamente la trayectoria de las primeras fijaciones. 2. Sobre el mapa de ganancia de información: donde se toma el máximo para la siguiente fijación: agregar ruido en esta etapa podría alterar el valor máximo y, por lo tanto, modificar la ubicación de la siguiente fijación. 3. Posición de llegada de movimientos oculares: Al elegir una fijación, se determina con probabilidad 50/50 si el modelo se dirige a ese lugar o a alguno inmediatamente cercano. A diferencia de los otros dos métodos, este no agrega ruido de los anteriormente mencionados a ningún componente del modelo.

Para los dos primeros componentes se utilizó ruido blanco como línea de base y dos distribuciones consideradas como más "naturales" a la hora de modelar procesos cognitivos: ruido rojo, generado utilizando ruido blanco (Zhivomirov, 2018) y ruido de Perlin (Perlin 1985, 2002) (Fig. 1B-D). Las diferencias entre ellos residen en la intensidad del mismo dentro de las distintas frecuencias. Se utilizaron 3 valores iniciales del signal-to-noise ratio (SNR): 0, 10 y 25 dB, que corresponden a un valor alto, medio y bajo de ruido. El agregado de ruido se realiza de forma aditiva, generando el ruido y luego sumando sobre el prior o el mapa de información.

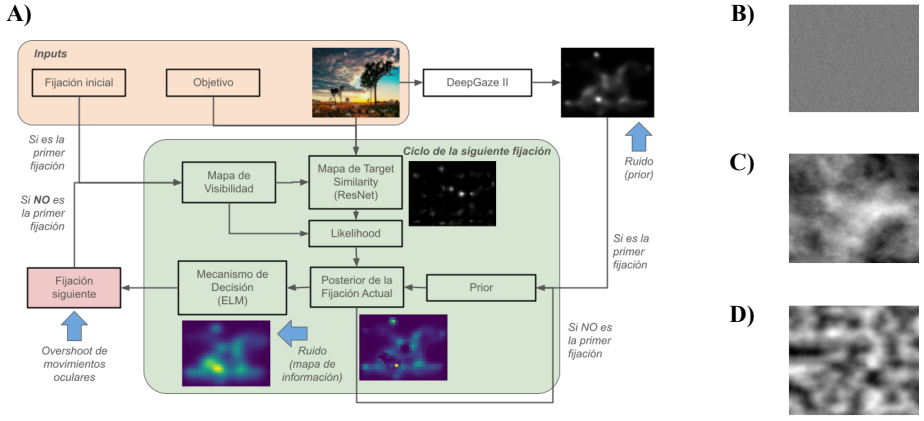

Fig 1. **A.** El modelo recibe como entrada la imagen, el objetivo y la primera fijación. Al principio calcula un prior inicial utilizando la red neuronal DeepGaze II (modelo que genera mapas de saliencia, entrenado a partir de tareas de exploración libre). Luego, computa iterativamente un likelihood que es, en pocas palabras, la similitud entre el target y la imagen circunscrita al campo visual actual. La siguiente fijación es aquella que maximizaría la información ganada en base a la posterior y a la visibilidad. Las flechas azules indican los componentes del modelo a los que se les añade ruido. **B.** Ruido blanco. **C.** Ruido rojo o marrón. **D.** Ruido de Perlin.

2.4 Métricas

Rendimiento acumulado. Se acumula la proporción de pruebas en las que se logró encontrar el objetivo en función del número de fijaciones realizadas. Se computa el área debajo la curva (*AUC*) para resumir los resultados. Se reporta el desvío estándar hasta la cuarta fijación (*stdF4*) y hasta la séptima fijación (*stdF7Acc*).
Multimatch (*MM*). Se representan los scanpaths como una secuencia de vectores en un espacio bidimensional (sacadas), donde la posición inicial y final de cada vector corresponden a fijaciones. Estas secuencias, que pueden tener diferente longitud, se comparan en cuatro dimensiones: forma (*MMvec*), largo (*MMlen*), posición (*MMpos*) y dirección (*MMdir*). Para cada imagen se comparan los scanpaths de los participantes entre sí y de las corridas de cada modelo entre sí. Luego, se promedia por imagen y se reportan los promedios y los desvíos estándar a lo largo de cada dimensión (*MMxxx_mean* y *MMxxx_std* respectivamente). Se reportan únicamente los resultados de los scanpaths de participantes y modelos que encontraron el target.

Scores. Se muestra un Score final junto a otros 2 scores que separan las métricas que miden la performance del modelo (*PerfScore*) y las métricas que miden la variabilidad del modelo (*VarScore*). En ambos casos acercarse a 0 implica valores cercanos a los obtenidos por los participantes. El *PerfScore* considera las métricas *AUCPerf* y *MMxxx_mean* mientras que el *VarScore* considera *stdF4*, stdF7Acc, y *MMxxx_std*. El *Score* considera todas las métricas

3 Resultados

Se muestran los resultados para 5 modelos (Tabla 1): un modelo determinístico como baseline (*elm_final*), y modelos con diferentes semillas para el cálculo de gaussianas (*elm_noise*), con ruido en el mapa de información (80db SNR) (*information*), con ruido en el prior inicial (25dB SNR) (*prior*), y un modelo que simula el overshooting de movimientos oculares (*overshoot*). Se utiliza el ruido rojo y el SNR que resulta mejor en cada caso, para *information* y *prior*.

Tabla 1. Resultados de métricas para diferentes modelos sobre el conjunto de datos utilizado para la experimentación. Se reportan resultados para 5 modelos en comparación a la base de los participantes.

Modelo	AUCperf	stdF4	stdF7Acc	MM_mean_avg	MM_std_avg	PerfScore	VarScore	Score
Humans	0.897	0.032	0.019	0.894	0.052	0.0	0.0	0.0
prior	0.744	0.019	0.011	0.897	0.053	-0.006	-0.296	-0.208
information	0.729	0.009	0.008	0.951	0.033	-0.024	-0.379	-0.28
overshoot	0.714	0.015	0.008	0.937	0.032	-0.018	-0.386	-0.283
elm_noise	0.729	0.011	0.007	0.958	0.030	-0.027	-0.409	-0.302
elm_final	0.721	0.001	0.001	1.0	0.0	-0.044	-0.845	-0.598

Al observar los valores del *Score*, se evidencia que el modelo *prior* fue el que obtuvo el valor más cercano a 0, lo cual indica que es el que mejor se aproxima al comportamiento humano en comparación con el modelo base *elm_noise*. La incorporación de ruido en los modelos *information* y *overshoot* también mejora su desempeño, aunque en menor medida que en el modelo *prior*.

4 Conclusiones

En este trabajo se analiza si el agregado de ruido sobre un modelo de búsqueda visual (ELM) puede aproximar la variabilidad existente entre humanos en la misma tarea, mientras que se mantiene la replicabilidad de resultados. Se experimentó con diferentes tipos y niveles de ruidos, en diferentes componentes del modelo: el prior inicial, el mapa de información generado antes de elegir cada fijación, y el punto de llegada de la sacada. Brevemente, se destaca la importancia del ruido (en particular ruido rojo) agregado en el prior inicial a la hora de aproximar variabilidad humana. Esto podría estar vinculado a los sesgos de los participantes a la hora de observar una imagen: dado que el prior inicial es un mapa de saliencia (el cual se genera para cada imagen una única vez), el ruido agregado sobre el prior puede estar modelando las diferencias en la saliencia para cada participante.

A futuro sería interesante realizar algunos experimentos de saliencia en humanos y medir efectivamente la variabilidad en la saliencia para utilizar como entrada al modelo. Además, en este manuscrito sólo se realizó una prueba de concepto de la variación en el punto de llegada de la sacada (overshooting) pero, dado que es un fenómeno conocido en la bibliografía, sería interesante profundizar en configuraciones más sofisticadas. Si bien se probó con distintos tipos de ruido, vale la pena revisar la literatura e intentar con otros tipos de ruido. Finalmente, se propone validar los resultados en otros conjuntos de datos (Travi et al., 2022), e integrar las medidas de variabilidad al mismo. Al mismo tiempo que sería importante extender a nuevas métricas, más allá del desvío estándar de las existentes, para comparar distribuciones de scanpaths.

4.1 Referencias

Bujia, G., Sclar, M., Vita, S., Solovey, G., & Kamienkowski, J.E. (2022). Modeling human visual search in natural scenes: A combined Bayesian searcher and saliency map approach. Front in Sys Neurosci, 16, 882315.

Bujia, G., Ruarte, G., Sclar, M., Solovey, G., & Kamienkowski, J. E. (2025). Uncertainty during visual search: Insights from a computational model and behavioral experiment in natural stimuli. bioRxiv, 2025-01.

Najemnik, J., & Geisler, W. S. (2009). Simple summation rule for optimal fixation selection in visual search. Vision Research, 49(10), 1286–1294.

Perlin, K. (1985). An image synthesizer. ACM Siggraph Computer Graphics, 19(3), 287-296.

Perlin, K. (2002). Improving noise. In Proceedings of the 29th annual conference on Computer graphics and interactive techniques (pp. 681-682).

Ruarte, G., Care, D., Bujia, G., Ison, M. J., & Kamienkowski, J. E. (2024). Integrating Ideal Bayesian Searcher and Neural Networks Models for Eye Movement Prediction in a Hybrid Search Task. bioRxiv, 2024-11.

Travi, F., Ruarte, G., Bujia, G., & Kamienkowski, J.E. (2022). ViSioNS: Visual search in natural scenes benchmark. Advances in Neural Information Processing Systems, 35, 11987–12000.

Wolfe, J. M., & Horowitz, T. S. (2008). Visual search. Scholarpedia, 3(7), 3325.

Zelinsky, G. J., Chen, Y., Ahn, S., Adeli, H., Yang, Z., Huang, L., Samaras, D., & Hoai, M. (2021). Predicting goal-directed attention control using inverse-reinforcement learning. Neurons, behavior, data analysis and theory, 2021, 10.51628/001c.22322.

Zhang, M., Feng, J., Ma, K. T., Lim, J. H., Zhao, Q., & Kreiman, G. (2018). Finding any Waldo with zero-shot invariant and efficient visual search. Nat Comms, 9(1), Article 3730.

Zhivomirov, H. (2018). A method for colored noise generation. Romanian journal of acoustics and vibration, 15(1), 14-19.