TF-IDF based Scene-Object Relations Correlate With Visual Attention

Pelin Çelikkol
University of Potsdam
Potsdam, Germany
acelikkol@uni-potsdam.de

Jochen Laubrock
University of Potsdam
Potsdam, Germany
laubrock@uni-potsdam.de

David Schlangen
University of Potsdam
Potsdam, Germany
david.schlangen@uni-potsdam.de

ABSTRACT

The relative contribution of bottom-up and top-down attentional guidance is a central topic in vision research. Whereas attention is guided bottom-up by low-level saliency, top-down guidance involves the viewer's knowledge and expectations accumulated throughout a lifetime. Here we explore the influence of high-level scene-object relations on viewing behavior. To assess top-down guidance, we score the relevance of linguistic object labels using methods from document analysis. Specifically, we computed the term frequency-inverse document frequency (TF-IDF), a statistic that reflects how important a term is to a document. We use object TF-IDF to measure how important a specific object is to a scene category and use these scores to predict eye movement distributions over scenes. Our results show that scene-specific objects are more likely to be fixated. Object TF-IDF had an effect partially independent of image saliency, suggesting that an object's relevance for a scene category affects attention during scene perception.

CCS CONCEPTS

• **Applied computing** → **Psychology**; • **Computing methodologies** → *Natural language processing*.

KEYWORDS

eye tracking, attention, scene perception, document analysis, statistics

ACM Reference Format:
Pelin Çelikkol, Jochen Laubrock, and David Schlangen. 2023. TF-IDF based Scene-Object Relations Correlate With Visual Attention. In *2023 Symposium on Eye Tracking Research and Applications (ETRA '23), May 30–June 02, 2023, Tubingen, Germany*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3588015.3588415

1 INTRODUCTION

We are exposed to a vast amount of information from our environment, but our visual system can only process a fragment of our visual surroundings at a given time. Due to architectural constraints of our foveated visual system, we face the concurrent tasks of analyzing the object at fixation while at the same time selecting the next object of interest from our low-acuity visual periphery. Visual attention in scene perception is prioritized toward particular regions to process the abundance of information efficiently. Our eyes repeatedly shift position to focus on regions of interest when looking at a scene [Yarbus 1967]. A fundamental step in understanding visual processing is to investigate the underlying mechanisms that produce the shifts in attention reflected by eye movements.

1.1 Bottom-up and top-down guidance of attention

Theories of visual attention disagree on the relative importance of low-level scene features [Itti and Koch 2001; Itti et al. 1998] and of the high-level meaning embedded in a scene [Bar 2004; Henderson 2007; Võ 2021] for visual selection. Theories of bottom-up visual saliency focus on salient contrasts in low-level perceptual features [Itti and Koch 2001]. Low-level features encompass the image-based bottom-up cues, such as luminance, color, intensity, and orientation. In this approach several low-level feature extraction processes operate in parallel, and attention is consequently focused on the most salient or conspicuous scene region while suppressing the previously attended locations. Whereas theories based on visual saliency provide an essential understanding of scene processing and offer biologically inspired computational models of visual attention [Itti and Koch 2001], theories focusing on top-down modulation of attention stress the importance of higher-level conceptual processes such as the viewer's high-level scene understanding, knowledge, and expectancies that bias activation in lower-level feature maps. Whereas bottom-up processing is mostly driven by visual saliency, higher-level scene understanding makes contact to linguistic representations. Therefore, the use of linguistic measures computed on scene descriptions may be useful to demonstrate top-down guidance of attention in scenes. Here we adapt one such measure, term frequency-inverse document frequency (TF-IDF), to show that attention is guided by relevance of objects for scene classification.

1.2 Meaning-based guidance of attention

A line of research associated with top-down control raises the question of whether the meaning embedded in a scene modulates visual attention [Ferreira and Rehrig 2019; Henderson and Hayes 2018]. Here, meaning-based guidance encompasses the characteristics of individual objects within a scene as well as how these objects are arranged and interact with each other [Ferreira and Rehrig 2019; Hayes and Henderson 2021; Henderson and Hayes 2018; Rehrig et al. 2022]. According to the meaning-based approach to visual attention, the knowledge regarding our visual surroundings is learned over time through interacting with our environment, and gaze control is guided by the viewer's stored information about a particular scene

ETRA '23, May 30–June 02, 2023, Tubingen, Germany

ACM ISBN 979-8-4007-0150-4/23/05.
https://doi.org/10.1145/3588015.3588415

category [Henderson 2007]. A scene's gist is extracted quickly after scene presentation [Potter 1975], which then activates stored information and expectancies in relation to the scene [Võ and Henderson 2010]. Scene-related expectancies modulate how we attend to, identify, and search for an object. In case of a violation of scene-related expectancies, inconsistent objects are processed more slowly and identified with more errors than scene-consistent objects [Biederman et al. 1982; Loftus and Mackworth 1978], suggesting that the activated contextual information of a scene regulates the cognitive processing of individual objects [Bar 2004].

1.3 Methods to quantify meaning-based guidance

Although scene-object relations have been shown to modulate attentional processing, studies that directly measure the effect of an object's contextual meaning on visual attention are scarce. Previous research aimed to address this question using meaning maps of real-world images as a measure of object informativeness [Henderson et al. 2019]. The meaning maps were constructed using ratings from human annotators that rated individual scene patches based on their meaningfulness, resulting in a topographical representation of the scene's informative parts. When the influence of the resulting meaning maps on attentional guidance was tested in comparison to saliency-based maps, meaning maps explained additional variance in gaze behavior.

Although meaning maps provide a unique way of measuring how the meaning of scenes affects eye movements, they rely on laborious human ratings. In a follow-up study, Hayes and Henderson [Hayes and Henderson 2021] used a vector-space model from computational linguistics to measure an object's relevance to a scene category. They scored the object labels based on the pair-wise cosine similarities among all objects in a scene and in relation to the scene category labels. They found that high similarity scores increased the likelihood of an object being attended, further suggesting a link between object meaning and attentional guidance. Rather than using human ratings of scene patches, this approach only needs annotated objects.

1.4 The present study

In the present study we extend the findings on attentional guidance by scene meaning. We propose a category-based object informativeness variant of TF-IDF as a new and easily computable measure to quantify the relevance of an object to a scene. To evaluate whether object relevance guides attention, we conducted an eye-tracking experiment using naturalistic indoor scenes and used TF-IDF to predict gaze fixations. TF-IDF is a statistic widely used in text analysis that indicates the relevance of a term in a document among a group of documents. TF-IDF weighs the frequency of a term within a document by the informativeness of the term across all documents, thereby favoring document-specific terms. We applied the TF-IDF logic to assess the importance of objects (terms) for a given scene category (document), using the ADE20k corpus [Zhou et al. 2019], which contains an extensive image database of indoor and outdoor scenes together with dense object annotations. The resulting scores served as a measure of the object relevance to a scene category, as they signify the object labels that are diagnostic to a scene (e.g., a *toy* in a *child's room*) while outweighing the importance of object labels that might be frequently encountered in an indoor scene (e.g., a *wall*). We tested the effect of TF-IDF scores on the allocation of visual attention represented by an object's probability of being fixated, controlling for object size and location.

2 METHODS

2.1 Participants

Twenty-three students from the University of Potsdam (9 male, mean age: 25.7 years, range: 19 to 38 years) participated in an eye tracking experiment in exchange for course credit or 10 Euros per hour of participation. Participants had normal or corrected-to-normal vision. They were naive as to the purpose of the experiment and fluent in English. Before starting the experiment, all participants provided written informed consent.

2.2 Material

Participants viewed 145 house-related indoor scenes that are a part of the Tell-me-more corpus [Ilinykh et al. 2019], which provides additional verbal annotations for a subset of the ADE20k images [Zhou et al. 2019]. The corpus consists of indoor images and description sequences obtained in an independent experiment, where participants were asked to produce a multi-sentence description of a scene, imagining that they were asked to "Tell me more" after each sentence. The resulting dataset of 4410 images contains five descriptive sentences for each scene in increasing detail as well as more generic captions for a subset of 411 images. The dataset contains visual scenes that elicit category-related object expectations in the viewer along with various possible compositions of individual objects [Ilinykh et al. 2019]. Additionally, the corpus provides pixel-level object annotations as they are a part of the ADE20k corpus [Zhou et al. 2019].

We applied various filters to the Tell-me-more corpus to obtain a subset appropriate for experimental presentation. Images had to have a minimum resolution of 760 x 1024 pixels suitable for eye tracking, generic captions that we could later use in our caption-matching task (see Procedure), and a minimum caption length to ensure caption quality. Finally, we sampled from image categories that have at least 50 examples. The resulting subset consisted of 145 images from 12 scene categories.

2.3 Apparatus

Eye-tracking data were recorded using the EyeLink 1000 system (SR Research Ltd., Ottawa, Canada) with a sampling rate of 1000 Hz in a tower mount setup. The eye tracker recorded the gaze position during binocular viewing. Visual stimuli were rescaled keeping the aspect ratio intact, and presented on a 27-inch iMac screen. Participants viewed the images from a 70 cm distance. The experimental presentation was controlled using MATLAB [MATLAB 2022] using the Psychophysics [Kleiner et al. 2007] and Eyelink [Cornelissen et al. 2002] Toolboxes.

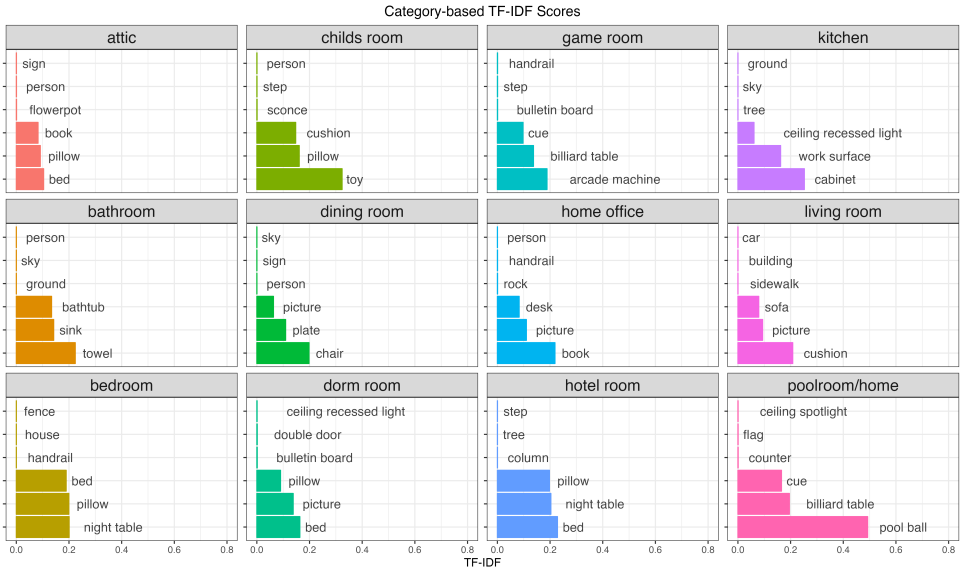

Figure 1: Illustration of TF-IDF as applied to objects in scenes. The three lowest- and highest-scoring object labels are shown for each of the 12 categories used in the experimental presentation.

2.4 Procedure

We presented each of the 145 images for 10 seconds to allow the participants to inspect the scenes in sufficient detail. A 9-point calibration was performed at the beginning of the experiment and after every 15 trials. Each trial started with a fixation cross presented at the center of the screen for 300 ms, followed by the presentation of the trial image. After the image presentation, participants were shown an image description, and their task was to indicate whether the description matched the previous image by pressing a key button. We balanced correct and incorrect match conditions, and correct and incorrect response alternatives were chosen from the same scene category. We chose the fairly demanding caption-matching task to ensure keeping the participant's attention throughout image presentation. Participants viewed each image once, resulting in 145 trials per participant and approximately 50 minutes of experiment duration. Presentation order was randomized between participants. Participants' eye movements were recorded throughout the experimental presentation.

2.5 Data preparation

2.5.1 Eye-tracking Data Preprocessing. We detected fixations and saccades using Eyelink's built-in algorithm that uses motion, velocity, and acceleration thresholds (0.1°, 30°/s, 8000°/s^2, respectively). We excluded a total of 37 trials (1.1%) due to poor calibration. This procedure resulted in 97685 fixations in total and an average of approximately 28 fixations per trial. We paired each fixation point with the corresponding object label obtained from the ADE20k corpus [Zhou et al. 2019]. When a fixation point corresponded to more than one bounding box, we took the smallest corresponding bounding box into account to ensure the precision of the fixated object.

2.5.2 Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF is a statistical method commonly used in information retrieval. It was introduced as a method to statistically assess term specifity in a collection of documents [Jones 1988]. While the TF-IDF score correlates with the term frequency in a document, it is outweighed by the document frequency in which the term appears, thus offsetting the importance of generic terms that frequently appear regardless of the document type. Given the term t in a document d, and D as a collection of documents, the term frequency (TF), inverse document frequency (IDF), and TF-IDF are calculated as follows, respectively:

$$TF(t,d) = \frac{freq(t,d)}{|d|} \tag{1}$$

$$IDF(t,D) = log(\frac{|D|}{|d \in D : t \in d|}) \tag{2}$$

$$TFIDF(t,d,D) = TF(t,d) \times IDF(t,D) \tag{3}$$

Here, $freq(t,d)$ denotes the number of times the term t appears in the document d, where $|d|$ is the total count of terms in the document. The total number of documents in the corpus is denoted with $|D|$, whereas $|d \in D : t \in d|$ indicates the number of documents with the term t in it.

We applied the TF-IDF method to the object labels obtained from the ADE20k corpus [Zhou et al. 2019]. In this way, we aimed to define a measure that represents the relevance of an object (e.g., *oven*) for a particular scene category (e.g., *kitchen*) while devaluating object labels that appear frequently regardless of the category (e.g., *wall*). We took the singular form of plural object labels and only used labels that indicate objects as a whole (e.g., *table*) and not object parts (e.g., *table leg*). We applied this analysis to the whole ADE20k corpus [Zhou et al. 2019]. The resulting data set contained 869 scene categories and 2270 object labels.

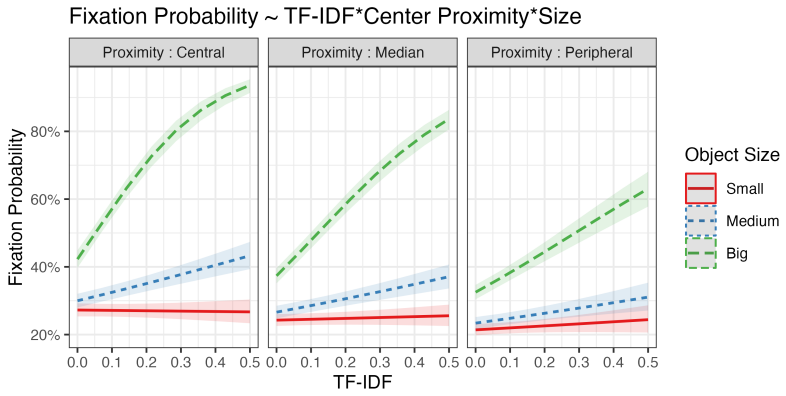

Figure 2: Three-way interaction of object TF-IDF, object size, and center proximity shown for the lower, median, and upper quartile values of the predictors. Predictors were only discretized for visualization purposes, they entered as continuous variables in the analysis.

To compute the TF-IDF score of object labels, we calculated the term frequency of an object label given a scene category, and weighted it by the number of scene categories the object label appears in. Figure 1 illustrates the measure. It depicts the three highest- and lowest-scored object labels for each scene category used in our experimental presentation. In general, high object TF-IDF scores indicate contextually coherent objects of a scene category (e.g., a *pool ball* in a *pool room*), whereas low object TF-IDF scores are associated with object labels that are not characteristic to a scene (e.g., *sky* in a *dining room*). We used the object TF-IDF scores as a predictor of gaze behavior.

2.6 Data analysis

We used a binomial generalized linear mixed model (GLMM) with a logit link function to analyze the effect of object TF-IDF scores on fixation probability, accounting for possible subject- or scene-based random effects. Models were fit using the 'lme4' package [Bates et al. 2015] in the R statistical computing environment [R Core Team 2022]. The binary dependent variable was an object's fixated/non-fixated status, and we added object TF-IDF as a fixed effect. To control for the well-known effects of center bias and object size on gaze behavior, we added these as fixed effect covariates. Interactions of the fixed effects were also analyzed. We fitted random intercepts for each subject and scene.

2.7 Image Saliency

To assess to what extent object relevance explains variance in eye movement distributions when low-level saliency is taken into account, we conducted an additional analysis using object saliency scores along with TF-IDF. We computed a saliency map for each image using the Itti and Koch image saliency model [Itti and Koch 2001] and scored each object based on the maximum saliency value within its bounding box area.

To assess the combined effect of object saliency and relevance on fixation distributions, as well as the unique effect of relevance, we fit three GLMMs: The baseline model only included the trivial covariates of object size and center proximity as fixed effects. We fit two additional models with the same structure, adding object saliency to the second model and additionally adding TF-IDF to the third model. As in the main analysis, scenes and subjects were added as random effects. We used likelihood ratio tests (LRT) and a comparison of the Bayesian Information criterion (BIC) among the three models to evaluate the goodness of fit. LRTs are generally used in model selection to compare nested models and to decide if certain predictors should be included. [1]

3 RESULTS

3.1 Task performance

Participants responded with an average of 77% correct when judging whether the image description matched the previous image or another image from the same category. This fairly weak performance suggests that image descriptions often do not narrow down an individual image within a category.

3.2 Fixation Probability

Results of the best performing model are summarized in Table 1 (Note that due to a large number of interaction effects, we excluded interaction effects of saliency scores for readability but added the complete results in the Appendix). Trivially, we found significant effects of object size and center proximity on the fixation probability: Larger objects were more likely to be fixated than smaller objects, and peripheral objects were less likely to be fixated than central objects. Object saliency scores significantly increased fixation probability. Most importantly, we found that higher TF-IDF scores significantly increased the probability of an object being fixated, considering all other factors. There were significant interaction effects among all variables; we illustrate the three-way interaction effects among the TF-IDF scores, object sizes, and center proximity in Figure 2. The higher the TF-IDF, the bigger was the size effect. Irrelevant big objects did not necessarily get fixated. The TF-IDF effect was stronger for more central objects, possibly indicating

[1] Data preprocessing and analysis code will be made available at: https://osf.io/ptury/?view_only=a6748dd9571d408798de6f0e4936c51b

Table 1: Results of the Generalized Linear Mixed-Effect Model

	Estimate	*SE*	*z*	*p*	*sig*
Intercept	−0.23	0.05	−4.88	< 0.001	***
TF-IDF	0.30	0.02	19.64	< 0.001	***
Saliency	0.13	0.01	11.25	< 0.001	***
Center Proximity	−0.43	0.01	−29.55	< 0.001	***
Object Size	1.67	0.02	72.54	< 0.001	***
TF-IDF * Center Proximity	−0.16	0.02	−8.01	< 0.001	***
TF-IDF * Object Size	0.54	0.03	16.61	< 0.001	***
Center Proximity * Object Size	−0.44	0.03	−14.42	< 0.001	***
TF-IDF * Center Proximity * Object Size	−0.32	0.04	−7.56	< 0.001	***

Table 2: Results of Likelihood Ratio Tests

Model	BIC	Chisq	Df	p
Baseline Model	95538			
Saliency Model	92086	3497.04	4	< 0.001
Full Model (Saliency + TF-IDF)	91573	604.27	8	< 0.001

limits of peripheral acuity for object identification. The three-way interaction indicates that both two-way interactions were mainly driven by relatively big objects (Examples of different object sizes and center proximity can be found in the Appendix).

Results of the analysis including saliency are summarized in Table 2. First, we found that the saliency model was a significantly better fit to the data when compared to the null model which included only the fixed effect covariates of object sizes and center proximity. Importantly, the full model, which additionally included object TF-IDF fit the eye movement distributions significantly better than the saliency model. BIC scores agree with LRTs in suggesting that the full model should be preferred, as evident by its lower BIC score.

4 DISCUSSION

We continuously gather information structures through repeated exposure to similar compositions within our environment and use the accumulated information to perceive and act upon the world efficiently. The knowledge and expectancies regarding our visual surroundings bring about a top-down modulation of visual processing following a rapid extraction of gist information. Building upon this approach, our study explored how scene-object relations affect fixation behavior. We showed that objects of high diagnostic value to a particular scene category are more likely to be attended. The effect of object relevance held when saliency was taken into account. This suggests a combined influence of object relevance and saliency on attentional allocation, where higher-level scene knowledge and expectancies drive attentional selection along with image-based conspicuous features. Our results support the notion that previously learned information about our environment guides attention toward the objects that are contextually coherent with that environment. Assessing contextual relevance might aid computational models of visual attention and lead to more accurate predictions of attentional allocation during scene viewing. Our study extends the findings regarding the role of scene semantics on visual perception (for a review, see [Henderson et al. 2019]) and offers an easy-to-compute method to assess certain object characteristics relative to a scene. The object TF-IDF measure is a potentially informative tool for any scene. Given that automated image annotation methods are readily available [He et al. 2017], our approach is potentially available for a wide variety of images. This opens up a new avenue to scene understanding.

Assessing what object labels repeatedly occur specific to a scene category can inform how likely an object will be encountered in that environment and how characteristic that object is to the scene, providing a useful measure to predict where attention will be directed. Our dataset consists of real-world indoor scenes where contextual inconsistencies and surprisal are unlikely. In the absence of a violation of the expectations, viewers tended to be more attentive to highly relevant scene regions in comparison to objects that were less informative about the classification of a scene. Besides the contextual informativeness of individual objects, typical object arrangements and locations constitute a scene's meaning. Previous work focused on the processing of spatially inconsistent objects in scenes and found that such objects are associated with slower attentional processing, suggesting a link between object arrangements and attentional guidance [Võ and Henderson 2009]. A text-based frequency approach similar to the present one might aid in assessing what objects are likely to co-occur in different scenes to test how these arrangements correlate with visual attention.

When assessing attentional guidance, the sequential nature of viewing is important in addition to the spatial distribution of attention, which we focused on in our current analysis. A scan path analysis in relation to scene meaning will further inform whether top-down influences facilitate attentional priority. The meaning-guided approach to scene understanding brings about further questions at the intersection of verbal and visual attention. We organize and put our thoughts in a certain order during language production,

thus linearizing our speech [Levelt 1989]. Linearization is essential when we describe a visual scene, where we need to incrementally process the visual details of a scene and verbalize them sequentially. Previous work on the linearization phenomenon showed that when speakers describe scenes, they visually attend to meaningful and informative parts, suggesting that visual and linguistic processing of scenes operate in coordination and are predominantly facilitated by meaning [Ferreira and Rehrig 2019]. The findings on multimodal processing entail a further exploration of how these cognitive modalities interact and to what extent top-down influences modulate these processes. Our experimental paradigm involved the reading of a scene description following an image presentation, thus, an in-depth analysis of viewing patterns during sentence reading might inform the nature of the attended expressions and their relation to the previously attended scene's characteristics.

Scene understanding is a complex process involving the viewer's generic knowledge about the world, expectancies, tasks, and goals. Building upon the present study of the scene meaning, we can better understand top-down guidance of human attention. In particular, we have shown that object distinctiveness, as measured by object TF-IDF, drives attentional selection. We are confident that further linguistics-inspired analyses of visual scenes can provide important insights into how we interpret and "read" visual scenes.

REFERENCES

Moshe Bar. 2004. Visual objects in context. *Nature Reviews Neuroscience* 5, 8 (2004), 617–629. https://doi.org/10.1038/nrn1476

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. https://doi.org/10.18637/jss.v067.i01

Irving Biederman, Robert J. Mezzanotte, and Jan C. Rabinowitz. 1982. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology* 14, 2 (1982), 143–177. https://doi.org/10.1016/0010-0285(82)90007-x

Frans W. Cornelissen, Enno M. Peters, and John Palmer. 2002. The Eyelink Toolbox: Eye tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods, Instruments, & Computers* 34, 4 (2002), 613–617. https://doi.org/10.3758/bf03195489

Fernanda Ferreira and Gwendolyn Rehrig. 2019. Linearisation during language production: evidence from scene meaning and saliency maps. *Language, Cognition and Neuroscience* 34, 9 (Jan. 2019), 1129–1139. https://doi.org/10.1080/23273798.2019.1566562

Taylor R. Hayes and John M. Henderson. 2021. Looking for semantic similarity: What a vector-space model of semantics can tell us about attention in real-world scenes. *Psychological Science* 32, 8 (Aug. 2021), 1262–1270. https://doi.org/10.1177/0956797621994768

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2980–2988. https://doi.org/10.1109/ICCV.2017.322

John M. Henderson. 2007. Regarding scenes. *Current directions in psychological science* 16, 4 (2007), 219–222. https://doi.org/10.1111/j.1467-8721.2007.00507.x

John M. Henderson and Taylor R. Hayes. 2018. Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision* 18, 6 (2018), 10–10. https://doi.org/10.1167/18.6.10

John M. Henderson, Taylor R. Hayes, Candace E. Peacock, and Gwendolyn Rehrig. 2019. Meaning and attentional guidance in scenes: A review of the meaning map approach. *Vision* 3, 2 (2019), 19. https://doi.org/10.3390/vision3020019

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Tell Me More: A Dataset of Visual Scene Description Sequences. In *Proceedings of the 12th International Conference on Natural Language Generation*. Association for Computational Linguistics, Tokyo, Japan, 152–157. https://doi.org/10.18653/v1/W19-8621

Laurent Itti and Christof Koch. 2001. Computational modelling of visual attention. *Nature reviews neuroscience* 2, 3 (2001), 194–203. https://doi.org/10.1038/35058500

Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence* 20, 11 (1998), 1254–1259. https://doi.org/10.1109/34.730558

Karen Sparck Jones. 1988. *A Statistical Interpretation of Term Specificity and Its Application in Retrieval.* Taylor Graham Publishing, GBR, 132–142.

Mario Kleiner, David H. Brainard, and Denis Pelli. 2007. What's new in Psychtoolbox-3? *Perception* 36 (2007), 1–16.

William J. M. Levelt. 1989. *Speaking: From Intention to Articulation.* MIT Press, Cambridge, MA.

Geoffrey R. Loftus and Norman H. Mackworth. 1978. Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human perception and performance* 4, 4 (1978), 565. https://doi.org/10.1037//0096-1523.4.4.565

MATLAB. 2022. *version 9.12.0 (R2022a).* The MathWorks Inc., Natick, Massachusetts.

Mary C. Potter. 1975. Meaning in visual search. *Science* 187, 4180 (1975), 965–966. https://doi.org/10.1126/science.1145183

R Core Team. 2022. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Gwendolyn Rehrig, Madison Barker, Candace E. Peacock, Taylor R. Hayes, John M. Henderson, and Fernanda Ferreira. 2022. Look at what I can do: Object affordances guide visual attention while speakers describe potential actions. *Attention, perception & psychophysics* 84 (2022), 1583–1610. https://doi.org/10.3758/s13414-022-02467-6

Melissa L-H. Võ. 2021. The meaning and structure of scenes. *Vision Research* 181 (2021), 10–20. https://doi.org/10.1016/j.visres.2020.11.003

Melissa L-H. Võ and John M. Henderson. 2009. Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of vision* 9, 3 (2009), 24–24. https://doi.org/10.1167/9.3.24

Melissa L-H. Võ and John M. Henderson. 2010. The time course of initial scene processing for eye movement guidance in natural scene search. *Journal of Vision* 10, 3 (2010), 14–14. https://doi.org/10.1167/10.3.14

Alfred L. Yarbus. 1967. Eye movements during perception of complex objects. In *Eye movements and vision.* Springer, Boston, MA, 171–211. https://doi.org/10.1007/978-1-4899-5379-7_8

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2019. Semantic understanding of scenes through the ADE20k dataset. *International Journal of Computer Vision* 127, 3 (2019), 302–321. https://doi.org/10.1007/s11263-018-1140-0