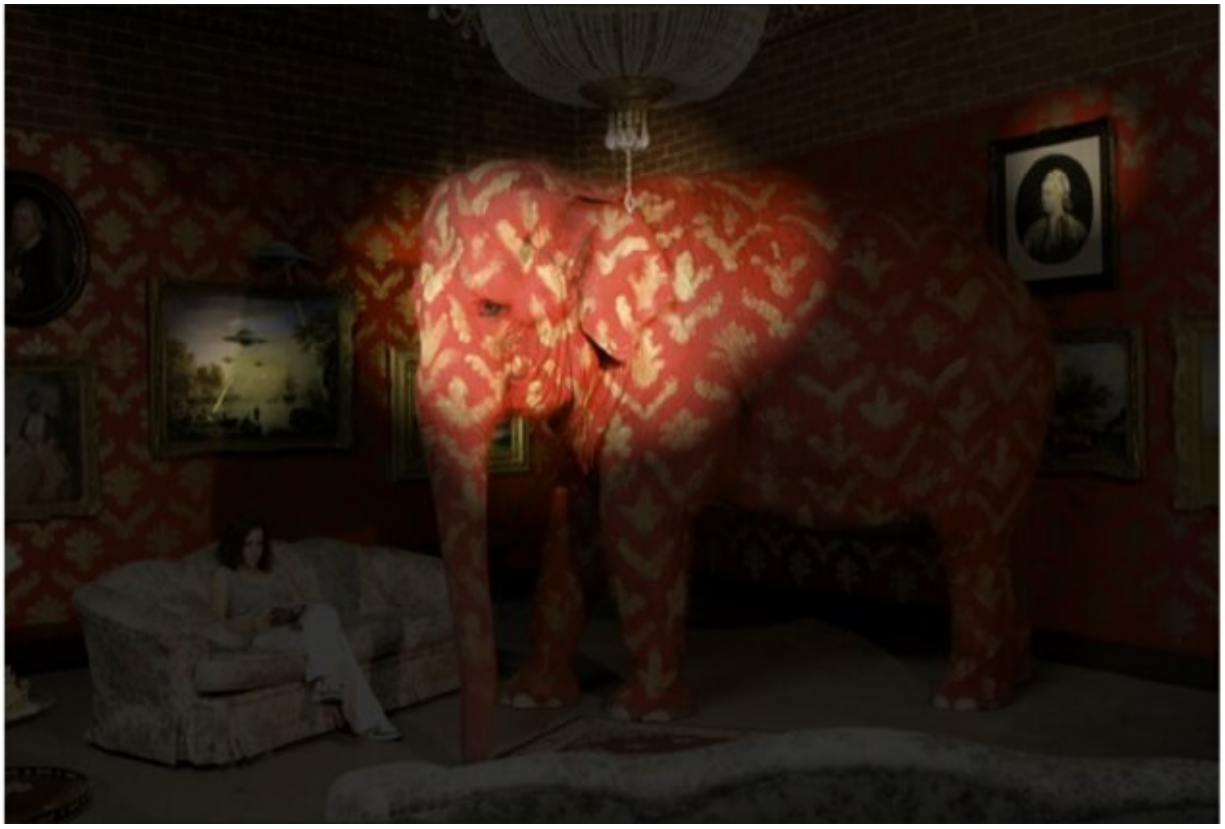
What jumps out in a photo changes the longer we look

June 18 2020, by Kim Martineau



An MIT study shows viewers' attention shifts the longer they gaze at an image. Given just a half-second to look at the photo at left, in online experiments, they focused on the elephant, as shown in this heat map. Credit: Massachusetts Institute of Technology

What seizes your attention at first glance might change with a closer

look. That elephant dressed in red wallpaper might initially grab your eye until your gaze moves to the woman on the living room couch and the surprising realization that the pair appear to be sharing a quiet moment together.

In a study being presented at the virtual Computer Vision and Pattern Recognition conference this week, researchers show that our attention moves in distinctive ways the longer we stare at an image, and that these viewing patterns can be replicated by artificial intelligence models. The work suggests immediate ways of improving how visual content is teased and eventually displayed online. For example, an automated cropping tool might zoom in on the elephant for a thumbnail preview or zoom out to include the intriguing details that become visible once a reader clicks on the story.

"In the real world, we look at the scenes around us and our attention also moves," says Anelise Newman, the study's co-lead author and a master's student at MIT. "What captures our interest over time varies." The study's senior authors are Zoya Bylinskii Ph.D. '18, a research scientist at Adobe Research, and Aude Oliva, co-director of the MIT Quest for Intelligence and a senior research scientist at MIT's Compute

What researchers know about saliency, and how humans perceive images, comes from experiments in which participants are shown pictures for a fixed period of time. But in the real world, human attention often shifts abruptly. To simulate this variability, the researchers used a crowdsourcing user interface called CodeCharts to show participants photos at three durations—half a second, three seconds, and five seconds—in a set of online experiments.

When the image disappeared, participants were asked to report where they had last looked by typing in a three-digit code on a gridded map corresponding to the image. In the end, the researchers were able to

gather heat maps of where in a given image participants had collectively focused their gaze at different moments in time.

At the split-second interval, viewers focused on faces or a visually dominant animal or object. By three seconds, their gaze had shifted to action-oriented features, like a dog on a leash, an archery target, or an airborne frisbee. At five seconds, their gaze either shot back, boomerang-like, to the main subject, or it lingered on the suggestive details.

"We were surprised at just how consistent these viewing patterns were at different durations," says the study's other lead author, Camilo Fosco, a Ph.D. student at MIT.

With real-world data in hand, the researchers next trained a deep learning model to predict the focal points of images it had never seen before, at different viewing durations. To reduce the size of their model, they included a recurrent module that works on compressed representations of the input image, mimicking the human gaze as it explores an image at varying durations. When tested, their model outperformed the state of the art at predicting saliency across viewing durations.

The model has potential applications for editing and rendering compressed images and even improving the accuracy of automated image captioning. In addition to guiding an editing tool to crop an image for shorter or longer viewing durations, it could prioritize which elements in a compressed image to render first for viewers. By clearing away the visual clutter in a scene, it could improve the overall accuracy of current photo-captioning techniques. It could also generate captions for images meant for split-second viewing only.

"The content that you consider most important depends on the time you have to look at it," says Bylinskii. "If you see the full image at once, you

may not have time to absorb it all."

As more images and videos are shared online, the need for better tools to find and make sense of relevant content is growing. Research on human attention offers insights for technologists. Just as computers and camera-equipped mobile phones helped create the data overload, they are also giving researchers new platforms for studying human attention and designing better tools to help us cut through the noise.

In a related study accepted to the ACM Conference on Human Factors in Computing Systems, researchers outline the relative benefits of four web-based user interfaces, including CodeCharts, for gathering human attention data at scale. All four tools capture attention without relying on traditional eye-tracking hardware in a lab, either by collecting self-reported gaze data, as CodeCharts does, or by recording where subjects click their mouse or zoom in on an image.

"There's no one-size-fits-all interface that works for all use cases, and our paper focuses on teasing apart these trade-offs," says Newman, lead author of the study.

By making it faster and cheaper to gather human attention data, the platforms may help to generate new knowledge on human vision and cognition. "The more we learn about how humans see and understand the world, the more we can build these insights into our AI tools to make them more useful," says Oliva.

More information: How Many Glances? Modeling Multi-duration Saliency: anelise.mit.edu/documents/md_s … rhm_camera_ready.pdf

TurkEyes: A Web-Based Toolbox for Crowdsourcing Attention Data: turkeyes.mit.edu/documents/TurkEyes.pdf

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology