This CVPR2015 extended abstract is the Open Access version, provided by the Computer Vision Foundation.

How many bits does it take for a stimulus to be salient?

Sayed Hossein Khatoonabadi[1], Nuno Vasconcelos[2], Ivan V. Bajić[3], Yufeng Shan[4]
[1,3] Simon Fraser University, Burnaby, BC, Canada. [2] University of California, San Diego, CA, USA. [4] Cisco Systems, Boxborough, MA, USA.

Visual attention mechanisms play an important role in the ability of biological vision to quickly parse complex scenes, as well as their robustness to scene clutter. A considerable research effort has been devoted during the past 25 years to the computational modeling of stimulus driven attention, typically through the development of models of visual saliency. Early approaches pursued a circuit driven view of the center-surround operation, modeling saliency as the result of center-surround filters and normalization [6]. Under these models, saliency is computed by a network of neurons, where a stimulus similar to its surround suppresses neural responses, resulting in low saliency, while a stimulus that differs from its surround is excitatory, leading to high saliency values. A particularly fruitful line of research has been to connect saliency to probabilistic inference. This draws on a long established view, in cognitive science, of the brain as a probabilistic inference engine [7], tuned to the visual statistics of the natural world [1]. In the cognitive science literature, it has long been proposed that the brain operates as a universal compression device, where each layer eliminates as much signal redundancy as possible from its input, while preserving all the information necessary for scene perception.

The compression-based models can be divided into two classes. The first class models saliency as a measure of stimulus information. For example, [2] advocate an information maximization view of visual attention, where the saliency of the stimulus at an image location is measured by the self-information [3] of that stimulus, under the distribution of feature responses throughout the visual field. If feature responses at the location have low probability under this distribution, self-information is high and the location considered salient. Otherwise, the stimulus is not salient. [5] proposes a similar idea, denoted Bayesian surprise, which equates saliency to the divergence between a prior feature distribution, collected from surround, and a posterior distribution, computed after observation of feature responses in the center. A second class of approaches equates saliency to a measure of signal compressibility. This consists of producing a compressed representation of the stimulus, through a principal component analysis [8], wavelets [9], or sparse decomposition [4], and measuring the error of stimulus reconstruction from this compressed representation. Incompressible image locations, which produce large reconstruction error, are then considered salient.

While many implementations of the compression principle have been proposed for saliency, none has really used a direct measure of compressibility. This has motivated us to investigate an alternative measure of saliency, directly tied to compression efficiency. The central idea is that there is no need to define new indirect measures of compressibility, since a direct measure is available at the output of any modern video compressor. In fact, due to the extraordinary amount of research in video compression over the last decades, modern video compression systems operate ever closer to the rate-distortion bounds. It follows that the number of bits produced by a modern video codec is a fairly accurate measure of the compressibility of the video being processed. In fact, because modern codecs work very hard to assign bits efficiently to different locations of the visual field, *the spatial distribution of bits can be seen as a saliency measure,* which directly implements the compressibility principle. Under this view, regions that require more bits to compress are more salient, while regions that require fewer bits are less.

We formalize this idea by proposing the *operational block description length (OBDL)* as a measure of saliency. The OBDL is the number of bits required to compress a given block of video data under a certain distortion criterion. This is a direct measure of stimulus compressibility, namely "how many bits it takes to compress." By leveraging extensive research in video compression, this is a far more accurate measure of compressibility than previous proposals, such as surprise, mutual information, or reconstruction error. The OBDL is equally easy to apply to images and video. For example, it does not require *weighting* the contributions of spatial and temporal errors, as the video encoder already uses motion estimation and compensation, and performs rate-distortion optimized bit assignments. Furthermore, because most modern cameras already contain an on-chip video compressor, it has trivial complexity for most computer vision applications. In fact, it is computed directly from the output of the entropy decoder, which is the first processing block in a video decoder.

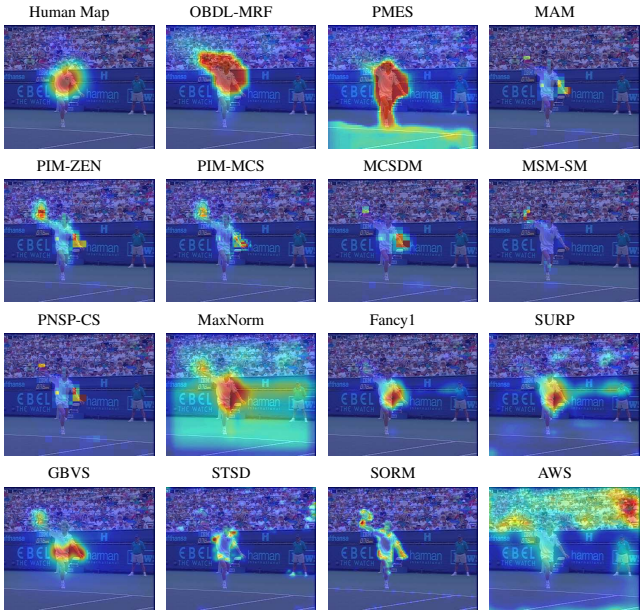

Figure 1: Saliency maps obtained by various algorithms on a video frame.

We show in this paper OBDL feature is highly predictive of eye fixations by comparing the statistics of the OBDL feature at human fixation points and non-attended locations, using two recent eye-tracking datasets. To account for global saliency effects, these are embedded in a Markov random field model. The resulting saliency measure is shown to achieve state-of-the-art accuracy for the prediction of fixations, at a very low computational cost (close to 30 fps with MATLAB implementation). Figure 1 illustrates the differences between the saliency predictions of various algorithms. Last but not least, the MATLAB code and data used in this study is available at www.sfu.ca/∼ibajic/software.html and www.svcl.ucsd.edu/publications.

[1] H. Barlow. Cerebral cortex as a model builder. In *Models of the Visual Cortex*, pages 37–46, 1985.

[2] N. Bruce and J. Tsotsos. Saliency based on information maximization. *Advances in Neural Information Processing Systems*, 18:155, 2006.

[3] T. M. Cover and J. A Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006.

[4] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. *Advances in Neural Information Processing Systems*, 21:681–688, 2008.

[5] L. Itti and P. F. Baldi. Bayesian surprise attracts human attention. *Advances in Neural Information Processing Systems*, 19:547–554, 2006.

[6] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998.

[7] D. Knill and W. Richards. *Perception as Bayesian Inference*. Cambridge Univ. Press, 1996.

[8] R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *Proc. IEEE CVPR'13*, pages 1139–1146, 2013.

[9] N. Sebe and M. S. Lew. Comparing salient point detectors. *Pattern Recognition Letters*, 24(1):89–96, 2003.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.