Part I
Keynote Lecture 1

VISUAL PSYCHOPHYSICS WITH NATURAL IMAGES

Isamu Motoyoshi
Department of Life Sciences, The University of Tokyo, Tokyo, Japan
<motoyosilab@gmail.com>

Humans can easily recognize objects, scenes, and faces in natural environments. The retinal image of 3D objects and scenes is highly complicated, and therefore many psychophysical studies have scrutinized the visual perception elementary features such as shape and color by using simplified stimuli on the assumption that analyzing such features is necessary for visual cognition in general. Psychophysical investigations, in concert with physiological evidence that neural sensors detect such images features, have certainly improved our understanding of the visual system. But can this assumption of lower-level elementary feature analysis be generalized to higher levels of visual cognition? For instance, does the attractiveness of an apple require the apple to be recognized, which in turn may necessitate the reconstruction of the apple's 3D shape and surface reflectance, which may itself imply precise analysis of curvature and gradients? Recent advances in computer vision and visual psychophysics with natural images cast a doubt on such a naïve hierarchical reconstruction scheme (e.g., Marr, 1982) and propose an alternative approach along the lines of 'task-specific vision' or 'direct perception' (e.g., Gibson, 1979).

Aristotle defined vision as knowing what is where by looking. To achieve this feat, the brain may not be obligated to reconstruct the physical properties (e.g., 3D geometry) of an external object. In principle, the brain could use only the information required to perform a specific task such as recognizing a person and judging her/his attractiveness. Of course, for a biological system, there is a premium on consuming less time and energy as greater efficiency may translate into better odds for natural selection.

One good strategy to achieve task-specific objectives is to use simple image features and their ensemble statistics which are already evident in the retinal image or in early visual cortex. Natural images are far from random—instead, natural images have a particular statistical structure (e.g., 1/f spectrum; e.g., Simoncelli and Olshausen, 2001). Moreover, at the statistical level, natural images often reflect the properties of external objects, scenes, materials, and so on. Thus, if a particular property is reflected in the statistical structure of the image, then the visual system may be capable of estimating that property from it directly. This short-cut strategy needs shallow computation and should therefore be very rapid. One may argue that image-feature representation is too poor for higher-order visual functions. However, recent evidence from psychophysics and computer vision reveals that the human visual system uses image statistics extensively to perform visual judgments on objects, scenes, materials, and their emotional values.

In the field of object recognition, for example, it has been suggested in the late 1990's that the visual system recognizes 3D object based on matching the 2D appearance of an object with canonical viewpoints stored in memory (e.g, Murase and Nayar, 1995). In the early 2000's, several computer vision studies have proposed powerful algorithms for object recognition based on populations of edge features (e.g., SIFT, SURF, and HOG; e.g., Lowe, 1999). These models have viewpoint-dependent characteristics consistent with physiological data from object-sensitive neurons in IT/TE (e.g., Poggio and Edelman, 1990; Logothetis et al., 1994).

In the field of scene perception, psychophysical studies have demonstrated that humans can categorize natural scenes with a latency of less than 100 ms (e.g., Thorpe

et al., 1996). Ultra-rapid categorization suggests that scene recognition depends at least partially on low-level feature representation in early visual cortex. Subsequent computer vision studies have supported this notion by showing that a simple model can extract the 'gist' of a scene from the distribution of image features (e.g., Oliva and Torralba, 2001).

The perception of surface properties such as glossiness and translucency has rarely been investigated because of its extreme complexity from a reconstruction-scheme standpoint. However, recent psychophysical and computational evidence shows that the human visual system can estimate surface properties based on very simple image statistics such as histogram skewness (e.g., Motoyoshi et al., 2007; Fleming, 2014). Recent studies further demonstrate that humans can judge the comfortableness/unpleasantness of a surface based on image statistics even faster than they recognize its material class (Motoyoshi and Mori, 2016).

All the aforementioned findings suggest a critical role for low-level image features in high-level visual cognition, including the judgment of emotional value. Of course, there is also significant amount of evidence that points to the limitation of image-based perception. For example, simple histogram statistics can explain (only) 75% of glossiness perception in daily objects (Wiebel et al., 2015). This indicates that, if needed, the visual system uses more elaborate information beyond image features. It seems that the visual system relies mostly on simple image features for quick and easy judgments and uses more information for detailed and accurate inspections. These two visual "modes" operate in parallel and are reminiscent of the classical human information-processing distinction of pre-attentive vs. attentive processes, and system 1 vs system 2 (e.g., Wolfe, 1998; Kahneman, 2003). To date, neither the neural representation nor the inherent computations involved in the latter process are understood. Future investigations may unveil them.

In addition to theoretical achievements, the series of studies cited here send us an important message with respect to psychophysical experiments: Experiments with natural or naturalistic images may lead us to the essence of the problem, whereas experiments using overly simplified stimuli—which are often designed based on a naïve understanding of the visual hierarchy—may mislead us. It should be always kept in mind that biological visual systems have evolved to interact efficiently with the natural world with limited resource and time.

Acknowledgements

This research was partially supported by JSPS KAKENHI JP15H05916 and JP15H03461.

References

Fleming, R. W. (2014). Visual perception of materials and their properties. *Vision Research,* 94, 62–75.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception.* Boston, MA: Houghton Mifflin.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review,* 1449–1475.

Logothetis, N., Pauls, J., Bülthoff, H., & Poggio, T. (1994). Shape representation in the inferior temporal cortex of monkeys. *Current Biology,* 4, 401–414.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. *The Proceedings of the Seventh IEEE International Conference on Computer Vision,* 1150–1157.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco, CA: Freeman.

Motoyoshi, I., Nishida, S., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature,* 447, 206–209.

Motoyoshi, I. & Mori, S. (2016). Image statistics and the affective responses to visual surfaces. *Journal of Vision,* 16(12):645.

Murase, H., & Nayar, S. K. (1995). Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision,* 14, 5–24.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision,* 42, 145–175.

Poggio, T., & Edelman, S. A. (1990). Network that learns to recognize 3D objects. *Nature,* 343, 263–266.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience,* 24, 1193–1216.

Thorpe, S., Fize, D., & Marlot., C. (1996). Speed of processing in the human visual system. *Nature,* 381, 520–522.

Wiebel, C. B., Toscani, M., & Gegenfurtner, K. R. (2015). Statistical correlates of perceived gloss in natural images. *Vision Research,* 115, 175–187.

Wolfe, J. M. (1998). Visual Search. In H. E. Pashler (Ed.), *Attention.* East Sussex, UK: Psychology Press.