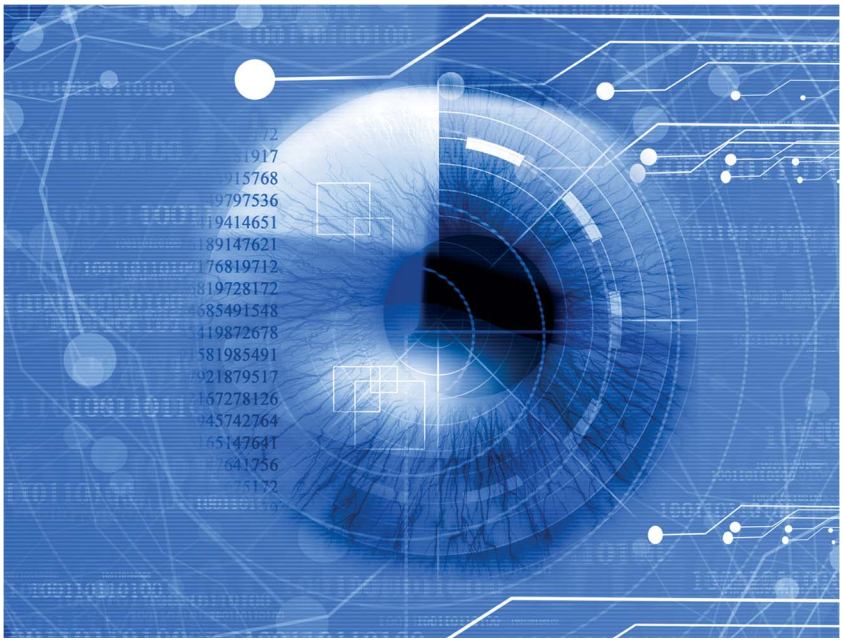
Perceptual Video Processing: Seeing the Future

By AL C. BOVIK

The University of Texas at Austin, Electrical and Computer Engineering, Austin, TX 78712-0240 USA



Over the last 30–50 years, video acquisition, processing, and display technologies have followed a remarkable trajectory. In 1980, I remember consumer video as analog television broadcasts that had just been widely deployed on cable television, freeing viewers from having only a few channel selections. Digital videos and video processing methods were largely laboratory research topics, with very few commercial products available. Today, the digital video revolution has been revealed as one of the most dramatic and transformative technological tides to sweep over human daily life. The analog television experience with its limiting resolution and noise has been replaced by crystal-clear, large-format all-digital visual experiences that have given a new meaning to the term "home theater." Large-format "high-definition" video presentations are now ubiquitous: we encounter them in the living room, the grocery store, the mall, the sports bar, the classroom—the list is endless. High-quality digital videos are also broadcast wirelessly to wherever we might be—in a convenient selection of sizes—ranging from small to not-so-small handheld devices (iPhone and Droid smartphones to iPads and laptops).

There is an overarching reason why digital video has become so pervasive, beyond just technical enablement. It is because we are largely visual creatures. Among humans that are not visually impaired, a large amount of the information that is received and processed by the brain is visual. Indeed, the human retina can transmit information at a rate of about 10 Mb/s to the brain, via more than 1 million nerve fibers comprising the optic nerve that passes information to the lateral geniculate nucleus (LGN) deep in the thalamus of the brain, where it is subsequently relayed to the primary visual cortex in the occipital lobe at the back of the head. There, on the order of half a billion cortical neurons actively decompose and process the visual data as it arrives, filtering and sorting quantitative space-time features of color, brightness, binocular disparity, and motion, creating sparse, high-information feature codes. These codes are sent to other brain areas that analyze and compute form, depths, velocities, and patterns, then send it on to further brain centers engaged in processes of navigation, recognition, attention, and decision. Just as the engineering field of video processing has changed remarkably over the past half-century, so too has vision science, as models, theories, and experiments on visual

Digital Object Identifier: 10.1109/JPROC.2010.2068371

function have gone far beyond "black-box" observations of visual behavior. There is now a wealth of deep and sophisticated knowledge on the functionality of the "visual brain," ranging from mappings of where certain types of processes occur, such as recognition, to detailed functional models of visual neurons, especially those conducting low-level "computational" types of processing that are amenable to engineering implementations. Admittedly, our current understanding of the visual brain is still in its infancy, with decades of inquiry ahead before it will be mapped. Yet, by 2010, a "standard model" of (at least) low-level visual processing has emerged that stands up to experiment, that is quantitative, and most importantly, that is accessible to, and useful to, video processing engineers.

Biologically motivated visual models have found their way into video processing systems. Indeed, when such models have been used, either directly in the design of video systems, or retroactively to explain or refine a successful video system design, they have often been of transformative significance. A good example was the choice of quantization weights used in the JPEG still image compression standard, based on human studies on the visibility of discrete cosine transform (DCT) basis functions [1]. The JPEG standard has been remarkably resilient, as it continues to dominate practice after nearly two decades since its first release. Another good example is the CIE XYZ color space, based on early color perception experiments by Wright [2] and Guild [3]. It underlies all of the color space representations used in color television standards since. Yet another clear example is the temporal contrast sensitivity function [4], which may be viewed as a "black-box" plot of the temporal frequency response of the overall human visual system (see Fig. 1). To the extent that the overall temporal visual system may be viewed as linear, the plot may be viewed as the approximate temporal passband of human visual response. Indeed, the upper

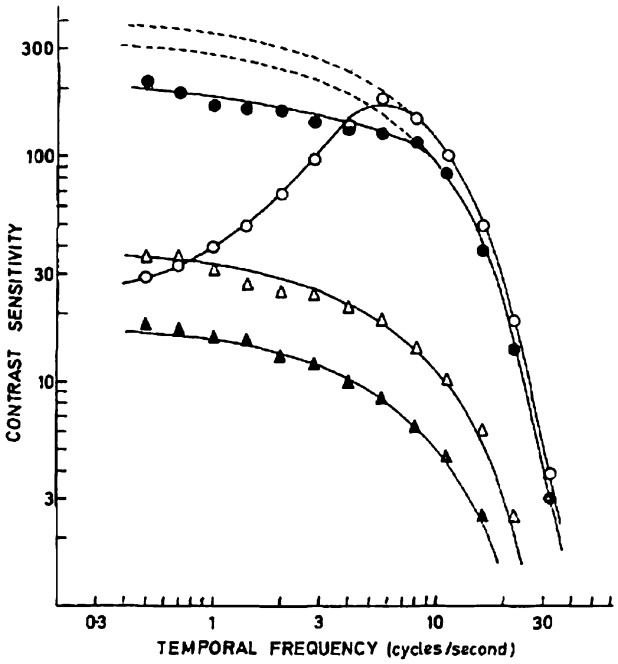


Fig. 1. *Temporal contrast sensitivity response of the human visual system. The curves differ by the spatial frequency of the temporal pattern presented: from uppermost to lowermost: 0.5, 4, 16, and 22 cycles per degree of viewing angle. From [4].*

cutoff frequency explains the temporal sampling frequencies or frame rates of movies, televisions, and monitors that have been designed to avoid visible "flicker," although early cinematographers probably did not think in these terms.

I. FROM NEURONS TO ALGORITHMS

The period 1960–1990 was a particularly rich one for discovery in the visual sciences. Examples of two (among others) that have profoundly affected digital image analysis methods are quantitative models of the responses of spatial-frequency, motion, and disparity-sensitive neurons in the visual cortex, and models of contrast masking of visual stimuli.

Seminal early work by Hubel and Wiesel [5] on the responses of cortical neurons indicated that they were sensitive to the orientations and widths of edges and bars in images. In the same period, Campbell and Robson [6] showed that the spatial visual passband may be divided into multiple orientation- and frequency-specific passbands, suggesting that visual processing proceeds early on in a manner akin to a spectrum analyzer. Further measurements of these "simple cell" responses led to quantitative linear-system models whereby, stated in engineering terms, images received in the cortex would be filtered by banks of spatially localized filters approximated as differences of 2-D Gaussian functions [7]. This model has been extensively used in the development of computer vision algorithms for analyzing images, e.g., influential theories for edge detection [8] and of scale- and rotation-invariant object recognition [9].

A refinement of the cortical simple cell model substituted 2-D extensions of Gabor's elementary functions [10] (Gaussian-modulated sinusoids), which possess optimized concentration in both space and frequency in the sense of the uncertainty principle [11]. This suggested that the visual system had evolved to optimally localize sensitivity to spatial image features while maximizing frequency

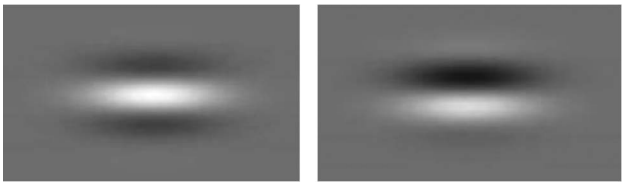


Fig. 2. *Quadrature pair of (left) cosine and (right) sine Gabor functions having a horizontal orientation.*

specificity. The population of simple cells covers a broad range of orientations and bandwidths (ranging from about 0.5 to 2.5 octaves). The Gabor model admits filters arranged in quadrature pairs (Fig. 2), allowing for convenient computation of their envelope responses. This model has been remarkably successful both as a description of early vision and as an image analysis model. It is used in some of the most popular and highly influential algorithms for image texture analysis [12], computational stereo [13], motion estimation [14], visual database search [15], iris [16] and face [17] recognition, and video quality assessment [18], to name just a few. The linear receptive field model of early human vision is a remarkable example of success in adapting human visual models to problems in image and video analysis. Indeed, it predated and seamlessly integrated with wavelet theory, which mathematically formalized many of the principles of 2-D multiscale analysis, 2-D subband decompositions, and orientation-sensitive processing already familiar to vision scientists. Indeed, some of the great early successes of wavelet theory were in image analysis [19] and in image compression [20].

Of course, human vision is hardly entirely linear. A good example of a nonlinear model that has proved successful in both vision science and in image and video processing is contrast masking [21]. Masking in vision is a phenomenon whereby one feature or aspect of a visual stimulus reduces or eliminates the visibility of another. Although there are many types of masking, a simple model for contrast masking is that the presence of local high-frequency energy in an image reduces the visibility of other high-frequency features.

An important video engineering application that benefits from this model is visual quality assessment, whereby the perceptual quality of an image or a video, as judged by an average human observer, is predicted by an algorithm [22]. In particular, flaws or impairments of the visual signal may be masked locally by strong high-frequency energy in the signal, thereby reducing the perceptual significance of the distortion(s)—and hence their effect on perceived quality. Fig. 3 nicely depicts contrast masking of JPEG image compression artifacts. In the image, JPEG blocking artifacts are pronounced on the girl's face and arms—which are smooth—and are far less visible in the surrounding flora. While these masking affects may be easily observed, quality assessment algorithms need to be able to quantitatively measure or predict this and other perception responses to distortions. As I will discuss below, algorithms that can reliably assess video quality in a manner that agrees with human subjective opinion are quite valuable in many practical scenarios, e.g., for comparing coding methods, measuring the video network quality of service, or as objective functions to optimize the production or communication of high-quality videos.

II. TODAY AND TOMORROW: STATISTICS AND PERCEPTUAL OPTIMIZATION

The use of quantitative visual models has, in the past, been more prevalent among computer vision experts than video processing engineers. This is not surprising since the goal of the former field is to create systems that can see, and the human visual system is a good model to attempt to emulate. Nevertheless, I believe that perceptual models will play an increasingly significant role in future optimizations of video acquisition, transmission, processing, and display systems. There are a number of compelling reasons why I take this view.

First, there is a developing explosion in video data delivery. Already, high-definition digital video systems are

Fig. 3. *Contrast masking of JPEG artifacts.*

being found in the home environment and elsewhere, with ever-increasing display sizes. On the wireless side, it is estimated that video cellular traffic will increase in volume by 50–100× times over the next five years, severely straining an already stressed wireless infrastructure capacity (exacerbated by iPhones and their relatives). The sudden popularity and interest in 3-D stereoscopic presentations, both in large-format and handheld displays, will significantly increase the need for capacity optimization going forward. This immense and growing volume of high-resolution video data introduces new challenges and the possibility of introducing new perceptual elements. Certainly, in the wireless realm, there is a great need to optimize video capacity; since there are limits on the degree to which traditional communication protocols can be stretched, I think that perceptually optimized networks represent the future of video networks. Indeed, I envision every switch, router, access point and base station, every set-top box, and cell phone being equipped with visual quality algorithms providing feedback to the overall distributed network control, thereby ensuring a uniform and optimized visual quality of service.

Second, I believe that visual models have a much larger role to play in optimizing video systems. This includes both existing models and emerging models. For example, with the onset of very large display sizes, the measured or predicted direction of gaze of viewer(s) is of great interest. There is a large and little exploited literature on modeling head and eye movements, on predicting the direction of gaze [23], and on exploiting video foveation to significantly conserve bandwidth [24]. I also believe that existing cortical models can be better exploited for many purposes. For example, the recently developed MOVIE index for video quality assessment [18] uses modern models of spatio–temporal visual decomposition in area V1 of visual cortex, of motion estimation and processing in area MT of the extra-striate cortex, coupled with spatio–temporal masking elements, to achieve outstanding quality prediction performance. As these models are refined and future models of higher level visual processing emerge, a more detailed and quantitative foundation for perceptual video processing and analysis will be made possible. Methods for probing visual function, such as functional magnetic resonance imaging (fMRI), are providing a window on the brain, making possible broad maps of the topography of visual function [25].

Third, approaches to visual modeling are becoming more deeply founded in statistics, owing to the realization that images and videos of the natural world follow statistical laws [26]. Biological vision systems have evolved and adapted to these statistical regularities; and image and video processing algorithms that are optimized under these statistical constraints can produce markedly improved results [27]. Statistical optimization with respect to the natural statistics of visual data is a nascent field that has begun to grow quite rapidly, with tremendous potential for video engineering experts to explore.

Fourth, in the last ten years, tremendous gains have been made in using perceptual quality prediction models to create algorithms that are able to assess the quality of images and videos in high accordance with human subjective judgments of quality [18], [22]. The corollary of this is that the existence of such algorithms implies that the broad spectrum of video acquisition, processing, communication, and display algorithms could be optimized with respect to predicted visual quality. However, while there is progress in this direction [28], and the payoff is likely to be large, the field is largely open. Traditional fidelity measures, such as the mean squared error (MSE) correlate poorly with perception [28], but are also easy to optimize against. Conversely, developing ways to optimize modern image and video quality assessment algorithms in practical scenarios will require solving significant problems in nonconvex and distributed optimization.

Despite these ongoing developments, I have found that university-level training in video processing largely fails to introduce the student to relevant aspects of vision science or to stress perceptual image processing. I think that video processing education should now stress relevant topics in vision science [29]. This would have the advantage of creating a more cross-disciplinary learning experience, greatly enriching the next generation of video engineers. Many universities offer courses in visual perception, commonly as part of a psychology curriculum. Failing that, or if the course is too specialized to be accessible to engineers, then perceptual material can be incorporated directly as part of image/video processing courses. In my view, this is the preferable route since it makes it possible to better integrate the perceptual material with video engineering algorithms, albeit requiring coverage of more material. My own courseware (and research) has been growing in this direction and the reader is invited to visit the LIVE SIVA website for free downloads [30]. I further believe that the tremendous growth in real-world digital video products and applications (and methods) implies that curricula in digital video should be expanded in any case. It is my hope, in fact, that the video engineer of the future will be part neuroscientist and behavioral psychologist. Indeed, one of my goals in writing this paper has been to provide a small overview of past innovations in this direction, as well as outlining my views on what is needed to create the perceptually optimized video viewing environment of the future. After all, the trend of increased digital visualization is only going to increase, and certainly within the not too-distant future our daily video experience will be not only 3-D, but increasingly immersive, blurring the lines between the experience of

reality and a synthesized projected reality. As long as we are going there, we may as well make the ride as comfortable, realistic, and of as high quality as possible. It is also my hope that future video engineers reading this will feel inspired to take on the fascinating cross-disciplinary challenges that lie ahead. ■

REFERENCES

[1] J. B. Pennebaker and J. L. Mitchell, *JPEG: Still Image Data Compression Standard*. New York: Springer-Verlag, 1992.

[2] W. D. Wright, "A re-determination of the trichromatic coefficients of the spectral colours," *Trans. Opt. Soc.*, vol. 30, pp. 141–164, 1928.

[3] J. Guild, "The colorimetric properties of the spectrum," *Phil. Trans. R. Soc. Lond. A* vol. 230, pp. 149–187, 1931.

[4] J. G. Robson, "Spatial and temporal contrast-sensitivity functions of the visual system," *J. Opt. Soc. Amer. A.*, vol. 56, pp. 1141–1142, 1966.

[5] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol. Lond.*, vol. 160, pp. 106–154, 1962.

[6] F. W. Campbell and J. G. Robson, "Applications of Fourier analysis to the visibility of gratings," *J. Physiol. Lond.* vol. 197, pp. 551–556, 1968.

[7] H. R. Wilson and S. C. Giese, "Threshold visibility of frequency gradient patterns," *Vis. Res.*, vol. 17, pp. 1177–1190, 1977.

[8] D. C. Marr and E. Hildreth, "Theory of edge detection," *Proc. R. Soc. Lond. B*, vol. 207, pp. 187–217, 1980.

[9] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 2, no. 60, pp. 91–110, 2004.

[10] D. Gabor, "Theory of communication," *J. Inst. Electr. Eng.*, vol. 93, pp. 429–457, 1946.

[11] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Amer.* vol. 2, pp. 1160–1169, 1985.

[12] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 1, pp. 55–73, Jan. 1990.

[13] D. J. Fleet and A. D. Jepson, "Computation of component image velocity from local phase information," *Int. J. Comput. Vis.*, vol. 5, pp. 77–104, 1990.

[14] D. J. Fleet, A. D. Jepson, and M. R. M. Jenkin, "Phase-based disparity measurement," *CVGIP: Image Understand.*, vol. 53, no. 2, pp. 198–210, 1991.

[15] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 837–842, Aug. 1996.

[16] J. G. Daugman, "High confidence visual recognition of persons by a test of statistical independence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 11, pp. 1148–1161, Nov. 1993.

[17] L. Wiskott, J.-M. Fellous, and C. Von Der Malsburgc, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, pp. 775–779, 1997.

[18] K. Seshadrinathan and A. C. Bovik, "Motion-tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[19] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[20] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. Image Process.* vol. 1, no. 2, pp. 205–220, Apr. 1992.

[21] G. E. Legge and J. M. Foley, "Contrast masking in human vision," *J. Opt. Soc. Amer.*, vol. 70, no. 12, pp. 1458–1471, 1980.

[22] Z. Wang, A. C. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[23] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[24] W. S. Geisler and J. S. Perry, "A real-time foveated multi-resolution system for low-bandwidth video communication," *Proc. SPIE—Int. Soc. Opt. Eng.*, vol. 3299, pp. 294–305, 1998.

[25] S. A. Engel, G. H. Glover, and B. A. Wandell, "Retinotopic organization in human visual cortex and the spatial precision of functional MRI," *Cerebral Cortex*, vol. 7, no. 2, pp. 181–192, 1997.

[26] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Amer.*, vol. 412, pp. 2370–2393, 1987.

[27] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.

[28] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it?—A new look at fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[29] A. C. Bovik, "What you see is what you learn," *IEEE Signal Process. Mag.*, Sep. 2010.

[30] U. Rajashekar, G. C. Panayi, F. P. Baumgartner, and A. C. Bovik, *SIVA—The Signal, Image and Video Audiovisual Demonstration Gallery*, Jan. 2002. [Online]. Available: http://live.ece.utexas.edu/class/siva/