DISCUSSION ARTICLE

Discussion

Ying Nian WU and Song Chun ZHU

VISION AND THE ART OF DATA AUGMENTATION

We have learned a lot from studying the sequence of artful works on EM/data augmentation authored by van Dyk and Meng. In this note, we discuss some of our thoughts (or rather speculations) on the problem of vision from the perspective of missing data modeling and data augmentation.

1. VISUAL COMPLEXITY AND THE MISSING DATA FRAMEWORK

When looking at our visual environment, we are not only aware of the rich details of the 3-D visual scene, but we also summarize the details into a simple description of "what is where," which provides the crucial information of a visual scene. Therefore, we can formulate the problem of vision in terms of the following three variables. (1) *Image*: a 2-D matrix (or a pair of matrix sequences). (2) *Details*: a representation of the 3-D scene in full detail. (3) *Summary*: the abstract description of "what is where." Of course, summary and details are relative concepts, and the two variables Details and Summary should be understood as the bottom and the top of a pyramid of increasingly abstract layers of visual concepts. We call this pyramid the "scene description." For instance, for a scenery image, the Summary may consist of abstract concepts such as river, trees, and their overall shapes, and Details may consist of concepts like water ripples and waves, tree leaves, and branches, and their shapes and locations.

The meaning of the Summary can be defined in terms of how the details look and how they are composed together. Mathematically, this amounts to a generative model $P(\text{Details} \mid \text{Summary})$, which decomposes the complexity in Details into deterministic redundancy

Ying Nian Wu is Assistant Professor, Department of Statistics, 8130 Math Sciences Building, University of California–Los Angeles, Los Angeles, CA 90095-1554 (E-mail: ywu@stat.ucla.edu). Song Chun Zhu is Assistant Professor, Department of Computer and Information Sciences, The Ohio State University, 2015 Neil Avenue, Columbus, OH 43210 (E-mail: szhu@cis.ohio-state.edu).

Journal of Computational and Graphical Statistics, Volume 10, Number 1, Pages 90–93

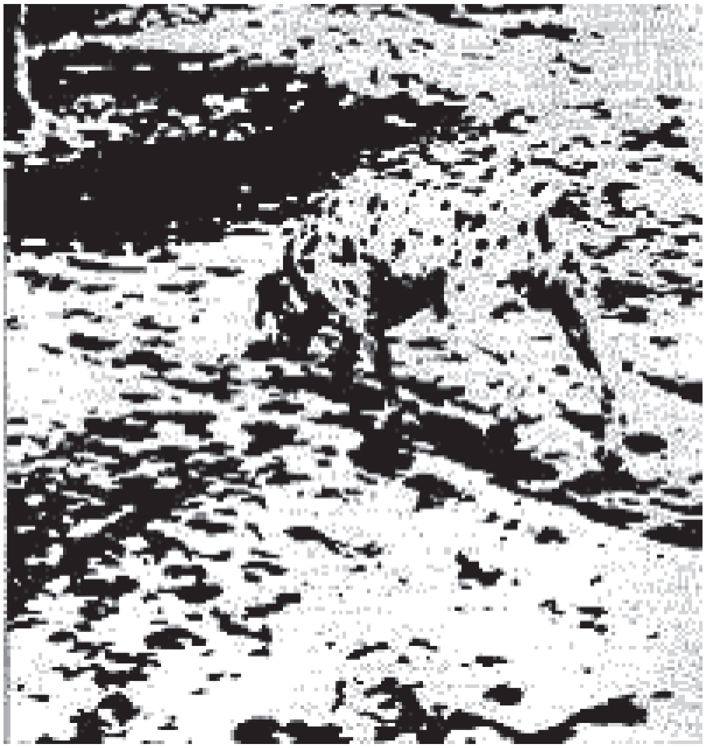

Figure 1. A Dog Walking on Snow. Purely bottom-up computation cannot find the contour of the dog.

and irrelevant randomness. Human vision perpetually summarizes complex details into simple patterns. With the detailed knowledge of the 3-D scene, the image can be rendered via Image = Graphics(Details). A more general form for this part of the model is P(Image | Details). Our prior knowledge on Summary can be represented by a distribution P(Summary). With this top-down generative model Summary → Details → Image, visual perception can be considered a process of computing the conditional distribution P(Summary, Details | Image).

This formulation clearly fits into the missing data framework, with Details being considered as the missing data. Then an EM/data augmentation algorithm can be derived as iterating the following two steps. (1) Scene reconstruction: imputing Details $\sim P$(Details | Image, Summary). (2) Scene understanding: abstracting Summary $\sim P$(Summary | Details).

Previous thinking on visual perception was often along the direction of bottom-up computation: Image → Details → Summary. This can be inadequate for visual perception because we sometimes need high-level knowledge to resolve uncertainties in perceiving low-level details. For example, in Figure 1, no matter how good our edge detector is, we cannot isolate the detailed contour of the dog without the help of the high-level knowledge. This point is reflected mathematically in the scene reconstruction step where we need to impute Details conditioning on both Image (bottom-up information) and Summary (top-down knowledge).

2. MENTAL OPTICS AND THE ART OF DATA AUGMENTATION

The scene description (Summary, Details) has both geometrical and photometrical aspects. The geometrical aspect includes the shapes, poses, and relative positions of the objects above a certain scale. It can be considered the "sketching" part of the scene description. The photometrical aspect includes lighting condition, reflectance properties of visible surfaces, as well as small-scale structures not describable in explicit geometrical terms. It can be considered the "painting" part of the scene description. So another way to look at the problem of vision is based on the three variables (Geometry, Photometry, Image). Estimating the Geometry from Image is crucial for our survival, and the estimated Geometry can be readily checked with the physical reality. Compared to the Geometry, the Photometry is only of secondary importance, and the introduction of Photometry may be viewed as an art of data augmentation, the purpose of which is mainly to assist the recovery of Geometry. For this reason, we should make Photometry and the augmented model P(Photometry) P(Image | Geometry, Photometry) as simple as possible, as long as the marginal P(Image | Geometry) leads to sufficiently accurate estimation of the Geometry. We call the mathematical representation of Photometry and the augmented model P(Photometry) P(Image | Geometry, Photometry) the "mental optics." There is no need for "mental optics" to go as deep as physical optics, because otherwise the modeling and computing can be made unnecessarily complicated without much gain. It is still largely a mystery how human brains perform this art of data augmentation. We need physics, psychology, and statistics to solve this puzzle.

Although the overall geometry provides the most important information of a visual scene, it is the complexity of the details and the photometrical aspect that defines perceptually realistic pictures. Therefore, understanding visual complexity and mental optics is crucial for visual perception and learning in computer vision and for realistic texturing and lighting in computer graphics.

3. CONCEPTUALIZATION AS DATA AUGMENTATION

For modeling images, one may argue that there are two major types of modeling strategies. One type consists of "exponential family models," which is based on the statistics of features; for example, responses from linear filters or edge detectors, which are computed deterministically from the observed image. The Markov random fields are models of this type (see, e.g., Wu, Zhu, and Liu 2000, and Zhu, Liu, and Wu 2000), and they are consistent with the bottom-up thinking in the research of visual perception. The other type consists of "data augmentation models," which introduce *hidden variables*; for example, linear basis, edges, bars, blobs, and so on as the causes for the observed image intensities. These hidden variables are to be imputed or inferred from the observed image. The models we discussed earlier are of this type and they are consistent with top-down thinking in the research of visual conception. In exponential family models, the data explain themselves (e.g., the Markov property of the Markov random fields), whereas in data augmentation models, the observed dependencies among the data are attributed to the sharing of common latent causes, and these latent causes become new concepts in our knowledge of data. For the purpose of conceptualization, the hidden causes should be independent so that they do not need further explanation, and at the same time, the image given the hidden causes should follow a simple model, so that the hidden causes provide a simple explanation for the dependencies among the data. If there are still remaining dependencies among the augmented latent variables, then we can further augment more abstract concepts; for example, lines, curves, flows, organizations, templates, and so on. This art of data augmentation or conceptualization may lead to a representational (instead of operational) theory of low-level vision, and may shed new light on Julesz's textons and Marr's primal sketches (see, e.g., Zhu and Guo 2000).

In some sense, our conceptualization of the world is an art of data augmentation. The data we continuously observe over time include images, sounds, touches, pleasure, pain, and our actions, and we want to make sense of the data—that is, to build a model, P(data), for our survival. For this purpose, our brains perform a data augmentation by introducing an extra variable, world, to simplify the modeling of the complicated dependencies among the sensory data. So we have an augmented model P(world) P(data $|$ world). In physics, people collect more data and find deeper laws, so the P(world) in physics becomes more profound, to the extent that the world and P(world) in quantum mechanics is so removed from the world and P(world) in our brains that we simply cannot imagine or conceive the quantum mechanical P(world) using our intuitive P(world).

ACKNOWLEDGMENTS

We thank the editor for kindly inviting us to contribute this discussion. The research is supported by NSF IIS-9877127 and NSF DMS-0072538.

REFERENCES

Wu, Y., Zhu, S. C., and Liu, X. (2000), "Equivalence of Julesz Ensembles and FRAME Models," *International Journal of Computer Vision*, 38, 245–261.

Zhu, S. C., and Guo, C. E. (2000), "Mathematical Modeling of Clutter: Descriptive vs. Generative Models," in *Proceedings of Spie Aerosense Conference On Automatic Target Recognition*, Orlando, FL.

Zhu, S. C., Liu, X., and Wu, Y. (2000), "Exploring Texture Ensembles by Efficient Markov Chain Monte Carlo—Towards a 'Trichromacy' Theory of Texture," *IEEE Pattern Analysis and Machine Intelligence*, 22, 554–569.