Topics in Cognitive Science 2 (2010) 217–225

ISSN: 1756-8757 print / 1756-8765 online
DOI: 10.1111/j.1756-8765.2009.01056.x

Visual Learning in Multisensory Environments

Robert A. Jacobs,[a] Ladan Shams[b]

[a]*Department of Brain and Cognitive Sciences, University of Rochester*
[b]*Department of Psychology, University of California, Los Angeles*

Received 24 February 2009; received in revised form 3 August 2009; accepted 16 August 2009

Abstract

We study the claim that multisensory environments are useful for visual learning because nonvisual percepts can be processed to produce error signals that people can use to adapt their visual systems. This hypothesis is motivated by a Bayesian network framework. The framework is useful because it ties together three observations that have appeared in the literature: (a) signals from nonvisual modalities can ''teach'' the visual system; (b) signals from nonvisual modalities can facilitate learning in the visual system; and (c) visual signals can become associated with (or be predicted by) signals from nonvisual modalities. Experimental data consistent with each of these observations are reviewed.

Keywords: Visual perception; Multisensory perception; Learning; Bayesian modeling

1. Introduction

The importance of using naturalistic environments to study human perception has been emphasized by many researchers (Findlay & Gilchrist, 2003; Hayhoe & Ballard, 2005; Land, 2003). Naturalistic environments contain complex sensory patterns, and our perceptual systems evolved and developed to allow us to sense and interpret these patterns. Until recently, it has been difficult for scientists to conduct carefully controlled experiments with naturalistic stimuli. However, new technologies, such as advanced computer graphics chips and virtual reality environments, are allowing researchers to investigate human perception in more realistic settings than has been possible in the past.

This trend is clearly evident in the study of visual learning. Many recent studies focus on how people improve at interpreting visual stimuli when these stimuli are part of multisensory

Correspondence should be sent to Robert A. Jacobs, Department of Brain & Cognitive Sciences, University of Rochester, Rochester, NY 14627. E-mail: robbie@bcs.rochester.edu

experimental environments. Because these environments are multisensory, they resemble naturalistic environments in important ways. Participants in these experiments can view objects, and they can often hear or touch these objects too. An important question that researchers are actively studying is how people take advantage of multisensory stimuli for the purposes of visual learning.

This article argues that multisensory environments are useful for visual learning because nonvisual percepts can be processed to produce feedback or error signals that the brain can use to adapt the visual system. To motivate this hypothesis, we describe a probabilistic framework, based on Bayesian networks, for thinking about visual learning in multisensory environments. This framework is useful because it ties together three important observations that have appeared in the scientific literature: (a) signals from nonvisual modalities can ''teach'' the visual system; (b) signals from nonvisual modalities can facilitate learning in the visual system; and (c) visual signals can become associated with (or be predicted by) signals from nonvisual modalities. In the next section, we describe the probabilistic framework. Following this, we describe experimental research consistent with the observations.

2. Bayesian network approach

We describe a probabilistic framework for thinking about visual learning in multisensory environments. This framework is based on a formalism known as Bayesian networks (Neapolitan, 2004; Pearl, 1988; Russell & Norvig, 2003). In general, a Bayesian network is a way of characterizing a joint probability distribution of several random variables. The network contains nodes, edges, and probability distributions. Each node corresponds to a variable. Each edge corresponds to a relationship between variables. Edges go from ''parent'' variables to ''child'' variables, thereby indicating that the values of the parent variables directly influence the values of the child variables. Each conditional probability distribution gives the probability of a child variable taking a particular value given the values of its parent variables. The joint probability distribution of all variables is equal to the product of the conditional probability distributions. For example, suppose that the joint distribution of variables A, B, C, D, E, F, and G can be factored as follows:

$$p(A, B, C, D, E, F, G) = p(A)p(B)p(C|A)p(D|A, B)p(E|B)p(F|C)p(G|D, E).$$

Then the Bayesian network in Fig. 1 represents this joint distribution.

Bayesian networks can represent the relationships between scene, feature, and input variables. Consider, for example, an environment with a coffee mug sitting on a desk, and an observer who views and touches the mug and desk. A schematic illustration of a Bayesian network appropriate for this situation is shown in Fig. 2. The nodes labeled ''scene variables'' describe the environmental scene. As a matter of notation, let S denote the scene variables. Based on the values of the scene variables, haptic feature variables, denoted F_{H}, and visual feature variables, denoted F_{V}, are assigned values. For instance, a coffee mug gives rise to both haptic features, such as curvature and smoothness, and visual features,

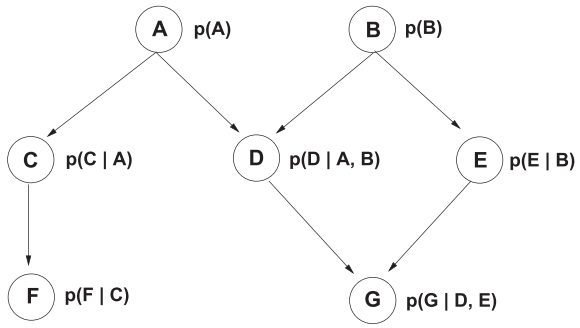

Fig. 1. An example of a Bayesian network.

such as curvature and color. The haptic features influence the values of the haptic input variables, denoted I_{H}, when the observer touches the mug. Similarly, the visual features influence the values of the visual input variables, denoted I_{V}, when the observer views the mug.

The input variables are ''visible'' because the observer obtains the values of these variables when he or she touches and views the scene. However, the feature and scene variables are not directly observable and are thus regarded as hidden or latent. The distribution of the latent variables must be computed by the observer from the values of the visible variables using Bayes' rule, an operation known as ''inference.'' For example, after touching and viewing a scene, the observer might want to infer the properties of the scene by calculating the conditional distribution of the scene variables given the values of the haptic and visual inputs:

$$p(S|I_{\mathrm{H}}, I_{\mathrm{V}}) = \iint p(S, F_{\mathrm{H}}, F_{\mathrm{V}}|I_{\mathrm{H}}, I_{\mathrm{V}})\mathrm{d}F_{\mathrm{H}}\mathrm{d}F_{\mathrm{V}}.$$

An advantage of Bayesian networks is that it is often computationally efficient to perform inference using a local message-passing algorithm (Pearl, 1988).

Visual perception is based on the posterior probability of the scene variables given the visual input variables $p(S \mid I_{\mathrm{V}})$ and therefore visual learning takes place when the observer adapts $p(S \mid I_{\mathrm{V}})$. To adapt this distribution, the observer needs feedback or error signals indicating how the distribution should be modified. How can the observer's visual system obtain these signals?

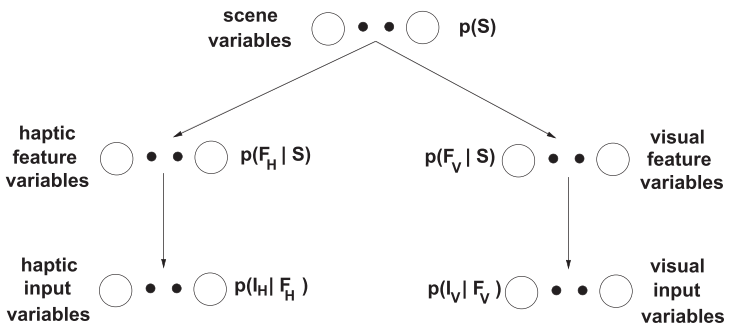

Fig. 2. A schematic of a Bayesian network appropriate for an observer that sees and touches its environment.

We consider three cases. In the first case, suppose that the observer touches and sees the coffee mug, and the observer has reason to believe that its haptic percept is highly reliable but its visual percept is highly unreliable (perhaps it is very noisy). The observer can infer the distribution of the scene variables in two ways: It can calculate the conditional distribution of these variables given the values of the haptic input variables $p(S \mid I_H)$ or given the values of the visual input variables $p(S \mid I_V)$. Because the observer regards its haptic percept as reliable and its visual percept as unreliable, it should regard $p(S \mid I_H)$ as more informative. Consequently, it can use $p(S \mid I_H)$ as a ''teaching'' signal and adapt its visual system (e.g., the distributions $p(F_V \mid I_V)$ and $p(S \mid F_V)$) so that $p(S \mid I_V)$ is closer to $p(S \mid I_H)$.

In the second case, the observer also touches and sees the coffee mug, but now it regards both its haptic and visual percepts as reliable. Again, the observer can infer the distribution of the scene variables in two ways: It can calculate the conditional distribution of these variables given the values of both haptic and visual input variables $p(S \mid I_H, I_V)$ or given the values of the visual input variables $p(S \mid I_V)$. The first distribution is based on more information, and thus it can be used as a teaching signal. That is, the observer can adapt its visual system so that $p(S \mid I_V)$ is closer to $p(S \mid I_H, I_V)$.

In the third case, consider an observer who touches the mug with his or her eyes closed or while looking at another object. In this case, the haptic input I_H can be used to predict the visual input I_V corresponding to the mug (via the distribution $p(I_V \mid I_H)$). Roughly speaking, it is as though the observer knows something about what the mug would look like based on the haptic percept. Of course, this prediction process can also occur when the visual input of the mug is available (i.e., while looking at the mug), and the difference between I_V and the predicted I_V can serve as an error signal used for visual learning. In addition to playing a role in learning, predicted visual quantities can be useful for other aspects of visual perception such as for guiding eye movements and attention, for generating top-down expectations about visual variables, and for visual imagery.

These three cases illustrate the main argument of this article, namely that multisensory environments are useful for visual learning because nonvisual percepts can be processed to produce feedback or error signals that people can use to adapt their visual systems. An important aspect of these cases is that they imply that observers know about the statistical relationships between different sensory modalities.

In summary, we have shown here that the Bayesian network framework is consistent with the ideas that signals from nonvisual modalities can ''teach'' the visual system (case 1), signals from nonvisual modalities can facilitate learning in the visual system (case 2), and signals from nonvisual modalities can be used to predict signals in the visual system (case 3). In the remainder of this article, we review experimental evidence from the scientific literature consistent with these ideas.

3. Nonvisual modalities teach the visual system

The idea that people learn how to visually perceive the world by comparing their visual percepts with percepts obtained from other modalities is an old one. Historically, it may

have been first proposed by Berkeley (1709/1910) who hypothesized that visual perception of depth results from associations between visual cues and sensations of touch and motor movements. A famous quote from Berkeley's book is ''touch educates vision.'' More recently, Piaget (1952) used similar ideas to explain how children learn to interpret and attach meaning to retinal images based on their motor interactions with physical objects. Empirical data supporting the general notion that motor interactions play a role in visual learning comes from prism adaptation studies in which subjects adapted to visual distortions produced by distorting lenses. Adaptation often occurs when subjects are allowed to interact with the environment (Held & Hein, 1958, 1963). In many studies, subjects only become aware of the visual distortion through their motor interactions (Welch, 1978).

Experimental data indicating that nonvisual modalities provide specific training signals that the visual system uses during learning has recently been obtained through the use of virtual reality environments. Atkins, Fiser, and Jacobs (2001) showed subjects vertically oriented cylinders whose horizontal cross-sections were either circular (the cylinder was equally deep as wide) or elliptical (a cylinder may have been either more deep than wide or less deep than wide). Visually, cylinders were defined by motion and texture cues. Importantly, a computer graphics ''trick'' was used that allowed independent control of the visual motion and texture cues. For example, a display may have contained a motion cue indicating that a cylinder was of one shape (e.g., circular cross-section), whereas the texture cue indicated that the same cylinder was of a slightly different shape (e.g., elliptical cross-section). In addition to seeing cylinders, subjects also grasped cylinders using a virtual reality force-feedback device. Half the subjects were initially trained in motion-relevant conditions—meaning that haptic and visual motion cues to shape were consistent, whereas the visual texture cue indicated an uncorrelated shape—and then trained in texture-relevant conditions—haptic and visual texture cues to shape were consistent, whereas the visual motion cue indicated an uncorrelated shape. The remaining subjects were trained in the reverse order. On visual test trials, it was found that subjects used the visual motion cue more after motion-relevant training than after texture-relevant training, and they used the visual texture cue more after texture-relevant training than after motion-relevant training. These data suggest that haptic information was used to determine which visual cue was more reliable, and thus which visual cue should be relied on more when making visual judgments of shape and depth. Other articles reporting experiments in which haptic percepts provided specific training signals used during visual learning include Adams, Graf, and Ernst (2004), Ernst, Banks, and Bülthoff (2000), and Atkins, Jacobs, and Knill (2003).

4. Signals from nonvisual modalities facilitate visual learning

The study discussed above shows that the consistency between visual cues and a nonvisual cue can serve as the basis for learning the relative reliabilities of visual cues. In this case, the cross-modal signal provides critical teaching input for the interpretation of conflicting visual cues. But visual learning can benefit from cross-modal information even when there is no conflict between visual cues. Recent studies have shown that learning of visual motion

detection and discrimination is facilitated when visual motion displays are accompanied by auditory motion during training (Kim, Seitz, & Shams, 2008; Seitz, Kim, & Shams, 2006). Although sound is not necessary for this kind of learning, it does nevertheless accelerate the course and increase the magnitude of learning when presented concurrently with visual stimuli during training. A group of observers was trained with random-dot kinematograms in a two-interval forced-choice (2-IFC) paradigm where the task was to detect the interval in which coherent motion was presented (Kim et al., 2008). Another group of participants was trained with exactly the same visual stimuli; however, the visual coherent motion was paired with auditory motion in the same direction (and the interval containing visual noise was paired with auditory noise). The two groups were tested on trials in which sound was completely absent. Despite the absence of sound during test trials, the group that was trained with auditory–visual stimuli learned faster and asymptoted at a higher level of performance. It is unlikely that this superior learning results from a higher arousal level during training due to the presence of sound because training with incongruent auditory–visual stimuli (moving in opposite directions) did not result in facilitated learning.

Although this study demonstrates facilitation of visual learning by sound on a low-level visual task, a similar type of facilitation has also been recently reported on a higher-level object recognition task (Lehmann & Murray, 2005; Murray et al., 2004). Subjects were presented with images of objects and were asked to report whether the image was novel (presented for the first time) or repeated (Murray et al., 2004). There were two presentations of each image in each block. The first presentation of some images was paired with the sound of the corresponding object (e.g., the image of a bell may have been paired with the sound of a bell), but the second presentation of all images was always in silence. Observers were more accurate at recognizing images of objects that were initially presented with sound compared to those that were initially presented without sound. This study shows that visual learning based on a single exposure to an image can benefit from cross-modal interactions.

5. Learning to predict visual signals from nonvisual signals

Associations between visual and nonvisual signals are ubiquitous in nature. When we see lightening, we expect to hear thunder; when a ball is about to hit the floor, we expect to hear and see a bounce; when we turn a light switch on, we expect to hear a click and see a change in brightness. These are all associations we have learned by experience. People are also capable of learning arbitrary cross-modal associations; for example, they can learn to read a foreign language text. However, the constraints, efficiency, and mechanisms of this type of learning are not well understood. A few recent studies have investigated this learning. In one study, observers were asked to perform an oddity detection task in visual, haptic, or visual–haptic stimuli in a 3-IFC paradigm (Ernst, 2007). During training, the luminance of the visual stimulus and the stiffness of the tactile stimulus were correlated. After training, observers' discrimination thresholds in the correlated trials and uncorrelated trials differed, indicating that they had learned the arbitrary correlation between luminance and stiffness.

The results of another recent study suggest that cross-modal associations can be learned even in the absence of a task. In this study (Seitz, Kim, van Wassenhove, & Shams, 2007), participants passively viewed a rapid stream of unfamiliar shapes concurrently presented with a rapid stream of unfamiliar sounds. Participants were not asked to perform a task and were not even aware that they were going to be tested afterwards. Unbeknownst to the participants, some statistical regularities (in the form of pairs and quartets) were embedded within the visual stream, within the auditory stream, as well as across the two streams during exposure. After the brief passive exposure phase, they were asked to judge the familiarity of stimulus ensembles (pairs or quartets) in a 2-IFC task. In addition to learning visual associations and auditory associations, the observers also learned auditory–visual associations, and these three types of learning appeared to be independent of each other.

These and other studies (Davies, Davies, & Bennett, 1982; Ernst, 2007; Howells, 1944) suggest that cross-modal associations can be learned efficiently and automatically. Once an association is learned, it can facilitate the detection and recognition of each sensory component through predictive or inference mechanisms (Friston, 2005; von Kriegstein & Giraud, 2006; Noppeney, Josephs, Hocking, Price, & Friston, 2008; den Ouden, Friston, Daw, McIntosh, & Stephan, 2009). For example, the sound of a bouncing ball would predict the location and time of the visual contact between ball and floor and, if that does not match the visual observation, it would result in an error signal that can then be used for visual learning (case 3 discussed above). If the visual and nonvisual modalities are calibrated with each other, visual learning can then be enhanced when the nonvisual modality teaches vision (case 1) or facilitates learning in vision (case 2).

6. Discussion

In our literature review, we identified three forms of visual learning that can benefit from nonvisual sensory stimulation: (a) cross-modal input can provide teaching signals for learning the relative reliabilities of visual cues when the cues provide conflicting information; (b) correlated cross-modal input can also accelerate and enhance visual learning for any single visual cue; and (c) through learned associations between visual and nonvisual signals, a richer multisensory representation is acquired, and this representation can enhance visual processing and learning. The classification of these phenomena into three categories does not imply that these forms of learning are independent of each other. To the contrary, we believe that these forms are highly interdependent and synergistic. For example, for (a) and (b) to occur, the statistical relationships between vision and one or more nonvisual modalities needs to be established first (i.e., [c] needs to take place). Furthermore, feedback provided by a nonvisual modality can only be interpreted correctly by the visual system (as in [b]) if processing of the two modalities is calibrated or in synch with each other (i.e., result of [a]).

Although these three forms of learning appear very different from each other, we argue that they all fit within the same Bayesian network framework. For all three classes of phenomena, a computational viewpoint suggests that cross-modal signals can be

processed to provide feedback and error information that can be used by the visual system for learning.

Evidence for multisensory learning has been accumulating in recent years, and the next step in this research program will be to understand the computational and neural mechanisms of the underlying learning processes. Although it is difficult to distinguish between mechanisms of learning and memory, research on visual learning and visual memory have, unfortunately, remained mostly segregated in the scientific literature. In discussing the facilitatory effects of sound on visual learning, we presented studies employing a visual perceptual learning paradigm (Kim et al., 2008; Seitz et al., 2006) and studies employing a memory task (Lehmann & Murray, 2005; Murray et al., 2004). This is an example of how studies from these two fields provide converging evidence for cross-modal facilitatory effects, and how findings of memory studies can inform learning research and vice versa.

It is not clear whether the multisensory benefits discussed here are specific to cross-modal stimuli or whether any additional sensory cue can provide similar benefits. For example, it is not known if more visual learning about a visual motion cue will take place when visual motion is correlated with a signal from another modality (e.g., an auditory signal) versus when it is correlated with another visual signal (e.g., a visual stereo signal). Our Bayesian network framework suggests that any signal that can provide additional feedback and error information can lead to the learning benefits discussed here, and thus these effects should generalize to any cue including within-modality cues. On the other hand, one could also imagine that cross-modal signals are special in that they are corrupted by independent noise processes, rendering them more informative than within-modality cues, which may be corrupted by dependent noise processes. Whether cross-modal benefits to learning are special is an empirical question that should be addressed by future research.

Acknowledgments

This work was supported by NSF research grant DRL-0817250 to the first author and by a grant from the UCLA Faculty Research Program to the second author.

References

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. *Nature Neuroscience*, *7*, 1057–1058.

Atkins, J. E., Fiser, J., & Jacobs, R. A. (2001). Experience-dependent visual cue integration based on consistencies between visual and haptic percepts. *Vision Research*, *41*, 449–461.

Atkins, J. E., Jacobs, R. A., & Knill, D. C. (2003). Experience-dependent visual cue recalibration based on discrepancies between visual and haptic percepts. *Vision Research*, *43*, 2603–2613.

Berkeley, G. (1709/1910). *Essay toward a new theory of vision*. London: Dutton.

Davies, P., Davies, G. L., & Bennett, S. (1982). An effective paradigm for conditioning visual perception in human subjects. *Perception*, *11*, 663–669.

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision*, *7*, 1–14.

Ernst, M. O., Banks, M. S., & Bülthoff, H. H. (2000). Touch can change visual slant perception. *Nature Neuroscience*, *3*, 69–73.

Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing*. Oxford, England: Oxford University Press.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *360*, 815–836.

Hayhoe, M., & Ballard, D. H. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*, 188–193.

Held, R., & Hein, A. (1958). Adaptation to disarranged eye-hand coordination contingent upon reafferent stimulation. *Perceptual and Motor Skills*, *8*, 87–90.

Held, R., & Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology*, *56*, 872–876.

Howells, T. H. (1944). The experimental development of color-tone synesthesia. *Journal of Experimental Psychology*, *34*, 87–103.

Kim, R. S., Seitz, A. R., & Shams, L. (2008). Benefits of stimulus congruency for multisensory facilitation of visual learning. *PLoS ONE*, *3*, e1532.

von Kriegstein, K., & Giraud, A. L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology*, *4*, e326.

Land, M. F. (2003). Eye movements in daily life. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (pp. 1357–1368). Cambridge, MA: MIT Press.

Lehmann, S., & Murray, M. M. (2005). The role of multisensory memories in unisensory object discrimination. *Cognitive Brain Research*, *24*, 326–334.

Murray, M. M., Michel, C. M., Grave de Peralta, R., Ortigue, S., Brunet, D., Gonzalez Andino, S., & Schnider, A. (2004). Rapid discrimination of visual and multisensory memories revealed by electrical neuroimaging. *Neuroimage*, *21*, 125–135.

Neapolitan, R. E. (2004). *Learning Bayesian networks*. Upper Saddle River, NJ: Pearson Prentice Hall.

Noppeney, U., Josephs, O., Hocking, J., Price, C. J., & Friston, K. (2008). The effect of prior visual information on recognition of speech and sounds. *Cerebral Cortex*, *18*, 598–609.

den Ouden, H., Friston, K., Daw, N., McIntosh, A. R., & Stephan, K. E. (2009). A dual role for prediction error in associative learning. *Cerebral Cortex*, *19*, 1175–1185.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Piaget, J. (1952). *The origins of intelligence in children*. New York: International Universities Press.

Russell, S., & Norvig, P. (2003). *Artificial intelligence: A modern approach*. Upper Saddle River, NJ: Prentice Hall.

Seitz, A. R., Kim, R., & Shams, L. (2006). Sound facilitates visual learning. *Current Biology*, *16*, 1422–1427.

Seitz, A. R., Kim, R., van Wassenhove, V., & Shams, L. (2007). Simultaneous and independent acquisition of multisensory and unisensory associations. *Perception*, *36*, 1445–1453.

Welch, R. B. (1978). *Perceptual modification: Adapting to altered sensory environments*. New York: Academic Press.