IRV: Learning to Integrate Visual Information Across Camera Movements

Peter N. Prokopowicz
Artificial Intelligence Laboratory
Department of Computer Science
University of Chicago
1100 East 58th Street
Chicago, Illinois 60637

Paul R. Cooper
Intelligent Perception and Action Lab
Dept. of Elec. Eng. and Computer Science
Northwestern University
1890 Maple Avenue
Evanston, Illinois 60201

Our eyes see well only what is directly in front of them; they must continually scan the faces, words, and objects around us. Perceptual integration is the process of combining the resulting jumpy, nonuniform images (Figure 1) into our stable, comprehensive perception of the world. Visual robots that scan their environment with moving cameras must also integrate visual information. We describe IRV, a visual robot that sees across camera movements [Prokopowicz, 1995]. Furthermore, IRV *learns* to integrate from experience, which consists of a series of random movements of a camera mounted on a motorized pan-tilt platform, observing the day-to-day activity in a laboratory. The learning procedure makes minimal assumptions, is robust, and scales. That is, learning proceeds without a prior analytic model, external calibration references, or a contrived environment, and can compensate for arbitrary imaging distortions, including lens aberrations, rotation of the camera about its viewing axis, and spatially-varying or even random sampling patterns.

IRV develops an accurate model of its own visual-motor geometry by learning to predict the sampled images that follow each random, but precise, camera movement (Figure 2). The model describes the relationship between any relative camera movement vector and the subsequent apparent motion of each pixel in the image. This relationship between corresponding pairs of visual points and camera movement vectors IS stored in a representation we call the *visual motor calibration map.* The map is filled over time from natural observations during development. Such table-based techniques for perceptual-motor development have been used to learn hand-eye coordination [Mel, 1990] and dynamic arm control policies [Atkeson, 1990]. For example, Mel's MURPHY memorized the relationship between the visual position of key points on its arm and the joint angles of the arm in that position. The individual experiences that IRV uses to fill its table are the visual shifts of pixels between successive images. This fundamentally ambiguous correspondence problem can not be determined from any single example. IRV overcomes this ambiguity by accumulating evidence from every repetition of each possible camera movement. Effectively, every *apparent* pixel correspondence (there are typically hundreds for every pixel) *votes* for the existence of an actual correspondence under the camera movement that just occurred.

Figure 1: The problem: A sequence of three overlapping views taken by a foveal, or spatially-varying, camera. Slight changes in viewpoint emphasize completely different details. Any understanding of the whole Bcene (lower right) demands integration of information across eye movements, both for human and foveal computer vision.

Eventually, enough votes accumulate to determine the true geometric relationship between pixels in successive images for the entire repertoire of movements, and the predictions become more accurate.

The calibration map is implemented as a connectionist visual memory that, during each movement, transforms visual information from the previous fixation into a reference frame centered on the new viewing direction. The visual shift of a single pixel for a particular movement is embodied as a three-way connection between units representing the camera movement, the visual location of a feature before the movement, and the location of the same feature after the movement. These connections are not present initially but *develop* during early

Figure 2: Learning to predict foveal images. Far left: Foveal image before five degree leftward camera movement. Left: Image of same scene after movement. Right: The ability to predict a post-movement image begins to develop after about 100 repetitions of the same relative movement. Far right: prediction improves after 1964 examples. Resolution in the center of the predicted image is limited by the original peripheral resolution.

Figure 3: Acquired visual-motor model allows calibration for arbitrary distortions. Top left: Another foveally-sampled image. Bottom left: the same image with arbitrary scrambling of the pixels. During learning, every input looks like this. Right: Same pixels, interpreted with calibration based on acquired visual-motor correspondences, after about 40,000 movements, or three days of learning.

experience by a hypothesize-and-test process. Over repeated practice movements, each visual unit notes any correlation between its inputs after the movement, and those of other units before the movement. The combinatorics of the problem make it impossible to record the frequency of every conceivable correspondence between pixels. Instead, only a small random set of possible correspondences is evaluated at any time. This reduces the number of connections needed for learning to $0(N_{picele}N_{movements})$, which is feasible both for artificial and biological systems [Prokopowicz, 1995].

The learned relationship between corresponding pairs of visual features and relative camera movements defines a motor-baaed metric that IRV uses to interpret an arbitrary non-uniform visual representation in terms of known movement angles. This interpretation assigns a true visual angle for each pixel, regardless of optical or sampling distortions. The calibration workB by constraining the angles assigned to pairs of corresponding pixels so that they are separated by the size and direction of the movement angle for which they have been found to correspond. As the accuracy of learned visual-motor correspondences improves, so does the accuracy of the constrained assignment of visual angles to pixels (Figure 3).

IRV can see over a field of view wider than that observable from the camera in a single position (Figure 1). IRV learns to visualize internally the location of peripheral image details that can no longer be resolved. The connectionist visual memory continually accumulates visual features near the fovea, and integrates them over time and eye movements by imagining where they would appear from the present viewing direction. We have replicated human psychophysical experiments which show that IRV can perceive and make accurate judgements about simple forms too large to fit in a single view.

The connectionist computational architecture and the experimental environment approximate the conditions of biological perceptual development; the learning algorithm is neurophysiologically plausible. Learning and mature performance both manifest time and space complexities commensurate with human abilities and resources. In other words, the *learning algorithm scales* completely. The results confirm the practicality of visual robots that learn to perceive the stability of the world despite eye movements, learn to integrate geometric features across fixations, and, in general, develop and calibrate accurate models of their own perceptual-motor systems.

References

[Atkeson, 1990] Christopher G. Atkeson Using local models to control movement. In *NIPS 3: Advances in Neural Information Processing Systems,* pages 316-323, edited by David S. Touretsky, Morgan Kaufmann.

[Mel, 1990] Bartlett W. Mel *Connectionist Robot Motion Planning: A neurally-inspired approach to visually-guided reaching.* Vol. 7. Perspectives in Artificial Intelligence, Edited by B. Chandrasekaran. Boston: Academic Press.

[Prokopowicz, 1995] Peter N. Prokopowicz. *The Development of Perceptual Integration Across Eye Movements In Visual Robots.* TR # 1 , Intelligent Perception And Action Lab, Institute for the Learning Sciences, Northwestern University, June, 1994.