Integrating Visual Information Across Camera Movements with a Visual-Motor Calibration Map

Peter N. Prokopowicz
Department of Computer Science
University of Chicago
Chicago, IL 60637
peterp@cs.uchicago.edu

Paul R. Cooper
Department of Computer Science
Northwestern University
Evanston, IL 60201
cooper@ils.nwu.edu

Abstract

Facing the competing demands for wider field of view and higher spatial resolution, computer vision will evolve toward greater use of foveal sensors and frequent camera movements. Integration of visual information across movements becomes a fundamental problem. We show that integration is possible using a biologically-inspired representation we call the visual-motor calibration map. The map is a memory-based model of the relationship between camera movements and corresponding pixel locations before and after any movement. The map constitutes a self-calibration that can compensate for non-uniform sampling, lens distortion, mechanical misalignments, and arbitrary pixel reordering. Integration takes place entirely in a retinotopic frame, using a short-term, predictive visual memory.

Introduction

The competing demands for wider field of view and higher spatial resolution suggest that computer vision systems will inevitably progress towards the trade-off that evolution selected for animal vision systems: foveal (or spatially varying) sampling combined with frequent camera movements. Thus, the integration of visual information across camera movements is a fundamental problem for lifelike computer vision systems. Figure 1 visually evokes the nature of the task.

A variety of possible solutions have been proposed, by researchers from psychology and neurophysiology as well as computer vision. These range from proposals that suggest that integration occurs in "retinotopic" coordinates, through theories that propose that integration occurs in a body- or head-based frame of reference, to theories that suggest integration occurs at a symbolic level of abstraction. Although the problem has received a great deal of attention, no completely convincing model has been developed.

We suggest that visual integration can be achieved through a representation we call the visual-motor calibration map, and that moreover such a map can be developmentally acquired. In this paper, we describe what constitutes a visual-motor calibration map, and how it can be used to solve problems of visual integration, including change detection, perceptual stability across eye movements, and the recognition of forms from multiple foveal observations. In particular, a developmentally-acquired map is sufficient to replicate psychophysical results on a recognition task involving eye movements. We describe the developmental process for the map elsewhere (Prokopowicz 1994).

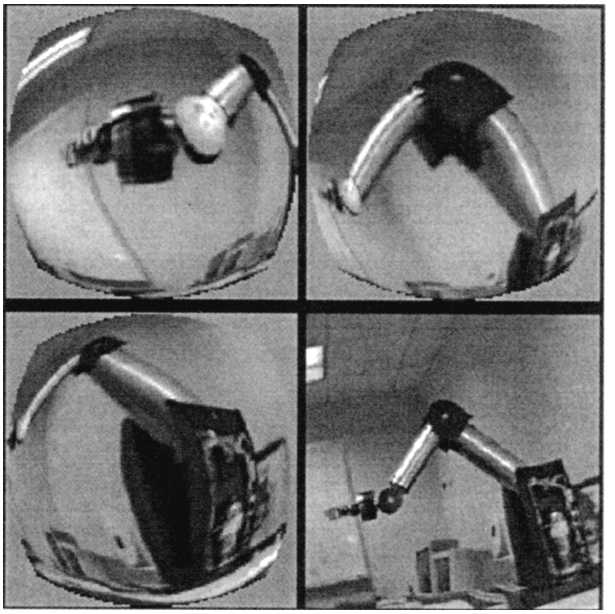

Figure 1: A series of overlapping views taken by a foveal, or spatially-varying, camera. Slight changes in viewpoint emphasize completely different details. Any visual understanding of the scene demands integration across fixations.

In brief, the visual-motor calibration map is a memory-based "table" representation of the relationship between camera movements and corresponding

pixel positions before and after a camera movement. Such a representation is neurobiologically plausible and scales-up to a realistically sized visual integration task. The map provides, in effect, a motor coordinate basis for visual information that allows visual tasks of extent larger than a single view to be computed. The map constitutes an adaptive visual-motor calibration that can compensate for effects including spatially-varying foveal sampling, random pixel reordering, and arbitrary (non-linear) lens distortions, all of which would be extraordinarily difficult to model analytically.

Visual Integration and Memory-based Calibration

It has long been known that our vision is many times more acute in the center of view than outside it, and that, to see, we unconsciously move our eyes two or three times every second, accumulating the details around us (Yarbus 1967). In seeking the mechanisms by which our perception can nevertheless be stable, and relatively uniform, theorists early on appreciated that disrupting the normal relationship between an intended eye movement and the viewpoint that would follow throws off these mechanisms: the world seems jumpy (Stratton 1897). Other experiments showed that, over several weeks, this effect can wear off, sometimes completely, and later rebound when the disruption is removed. These experiments suggest that in normal seeing, we unconsciously account for the expected visual effects of eye movements, and that this is a learned or adaptive ability.

Frames of reference in theories of perceptual integration One theory of perceptual integration holds that the stability and uniformity of perception corresponds to a stable and uniform internal visual representation, formulated in a head or body-centered frame of reference (Feldman 1985). Each succeeding fixation can be viewed as "painting" a different part of a larger inner image, one whose parts persist in memory long enough, perhaps a few seconds, to accumulate a detailed picture of the scene. This representation is invariant with respect to each eye movement, except that the immediate input is directed, or shifted, into the picture according to where the eyes currently point. This stable mental picture, consisting of pixels or some other primitive visual features, constitutes the effective input for the rest of perception, which goes on as if there were no small eye movements at all.

Others suggest that eye movements are not accounted for at this early a stage of perception, but at a much higher level of abstraction, in terms of the geometric relation between concrete objects or components (Pollatsek, Rayner, & Collins 1984). When a nose is perceived, for example, a representation of it is tagged with its visual location, taking into account the current direction of gaze. When the mouth is found, the distance between them can be inferred from the remembered position tags, even though they were perceived from slightly different viewpoints.

Usually, this tagging is thought to go on automatically for each item as it is recognized. A recent and more radical proposition holds that visual features are not compulsively memorized and tagged at any level of abstraction simply on the chance that they might later be usefully integrated with other features (O'Regan & Levy-Schoen 1983). Instead, integration across eye movements occurs only for those features that are part of a working model under construction as part of some specific perceptual task. When some unknown particular of the model is needed to continue with the task at hand, eye movements will be made toward where the information is likely to be found.

Physiologists have something to say about these hypotheses. So far, no invariant visual representations have been found; to the contrary, the visual system responds almost entirely to the retinotopic position of features (van Essen *et al.* 1991). However, it has been observed that in many areas of the visual system, activity shifts itself during an eye movement, as if to predict what the next retinotopic representation will be (Duhamel, Colby, & Goldberg 1992). This predicted activity can persist even without being reinforced with an actual input (Sparks & Porter 1983). In other words, features are envisioned retinotopically in a type of very short term visual memory. Also, it has been shown that subjects can often notice when something moves slightly during a saccade, even when they haven't been looking directly at it. Together, these experiments suggest that there could be a widespread, pre-attentive visual memory for perceptual integration that uses the current retinotopic frame of reference. This is the type of architecture we propose here.

Table-based perceptual motor control Regardless of which frame of reference and level of abstraction underlies integration across eye movements, the process must have access to an accurate quantitative model of the eye/head or camera/mount geometry. Without this, it would be impossible to understand the true geometric relationship between features or objects observed from different viewpoints. Traditional computer vision systems, which, to date, typically have not exploited moving cameras, rely on analytically determined models that are calibrated with reference standards. These models are neither easy to develop nor calibrate, especially considering the trend toward active vision and the potential long-term move towards

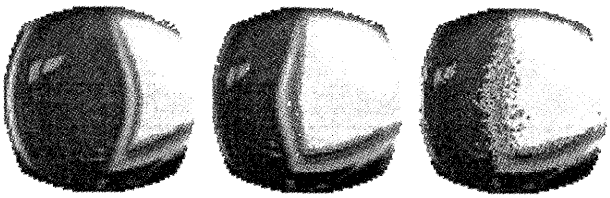

Figure 2: Foveal images before and after a camera movement are shown in the left and center columns. Predicted image, generated with an acquired visual-motor calibration map, is at right.

non-uniform sampling. Recently, there has been growing interest in solving visual problems without requiring traditional calibration (Faugeras 1992). Our work takes the approach that the necessary perceptual- motor models can and should be developed by the system itself from natural experience, and tuned continuously as part of every day activity.

As the representational basis for visual-motor calibration information, we propose a memory-based or table-based model of the relationship between camera movements and corresponding pixel locations before and after a movement. This representation is similar in spirit to the Mel's hand-eye kinematic models (Mel 1990) and Atkeson's arm-joint dynamic models (Atkeson 1990). In table-based control, the relationship between two components of a system are directly stored as tuples for every usable configuration of the system. Mel's system used the relationship between the joint angles of an articulated arm, and the retinotopic location of the arm's end-effector as viewed by a fixed camera. Intermediate factors, such as the arm's component lengths, or the camera's focal length, are not explicitly represented. Since it concerns only the direct relationship between relevant variables, such a model doesn't change form when intermediate components are added or replaced; it isn't even necessary to appreciate what these factors are.

The Visual Motor Calibration Map: a new representation for perceptual integration across eye movements

At its lowest level of abstraction, the problem of perceptual integration across eye movements is straightforward to state: *for any particular eye movement, what is the geometric relationship between a point A viewed before the movement and a point B viewed after?* This is the basic metrical information needed to combine or integrate visual information, in any form, from successive viewpoints. This relationship can be

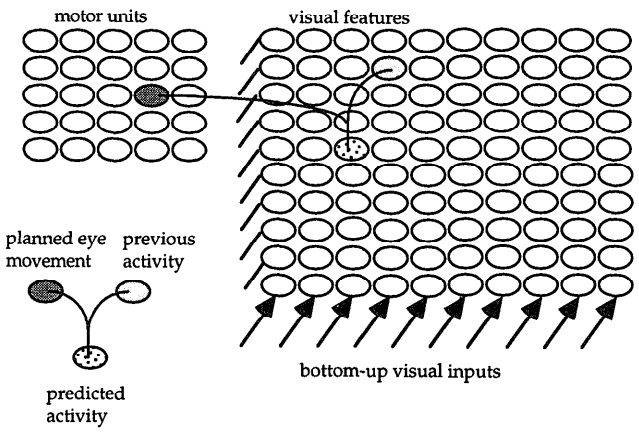

Figure 3: Connectionist architecture for predicting images across saccades. When a particular relative eye movement is about to be made, a motor unit produces a signal. A visual cell then predicts that its value will be that of the visual cell paired with this motor unit. After the movement, the cell receives bottom-up visual input from the sensor array.

described as tuple joining the two lines of sight defined by image points A and B, the visual angle V between the lines, and the eye-movement vector M. It is conceivable that this relation $R(A, B, M, V)$ could be determined and stored. The visual motor calibration map represents a similar relation that defines the image points A and B which correspond to the same line of sight($V = 0$), before and after a movement M.

To find the corresponding post-movement image location, given an eye movement and pre-movement location, the relation can be represented as a two-place look-up table. This has a natural connectionist equivalent (fig. 3) that is similar to other reference frame transformation networks (Feldman 1985). To find the third value of the relation given any other two, a slightly more complex network is needed (Feldman & Ballard 1982).

This relation, $R(A, B, M)$, can be used to predict an image following a particular camera movement M: for each pixel location A, copy that pixel's into location B if and only if $R(A, B, M)$ holds. In the same way, the map makes it possible to envision the post-movement location of any feature or object. We will also see that, if you know the pan and tilt angles of each movement M, the visual motor map completely calibrates the visual-motor system.

A look at the structure of IRV, our robotic visual motor system, will clarify the specific calibration prob-

lems that the map faces. The camera is a Panasonic GP-KS102, mounted on a computerized pan-tilt head, the Directed Perception model PTU. We restrict the head to 25 possible relative movements of roughly 2.5 to 5 degrees horizontally and/or vertically. The 400 by 400 digital image, spanning a visual angle of about 30 degrees, is resampled with an artificial foveal pattern (fig. 4) down to 80 by 80 non-uniform pixels, to produce the images like those in fig. 1. Finally, we randomly reorder the resampled pixels (fig. 7 top). An analytic visual model must account for the optical and sampling properties of the camera and artificial fovea, and the relation between the camera and its mount, which is complicated by the optical and mechanical axes not aligning. Arbitrary pixel ordering constitutes a completely general calibration problem, as well as a worst-case model of a developing organism in which the optic nerve and other fiber bundles scramble the topological ordering originally present on the retina.

As mentioned, a table-based model must fit in a reasonable space. We can restrict the visual-motor calibration relation so that for each camera movement, every image location before the movement corresponds to at most one image location after the movement; then, the map can be represented as a table with $N_{movements}N_{pixels}$ entries, 160,000 in IRV's case. A visual system scaled to roughly human levels would have about 1 million image locations, based on the size of the optic nerve, and 100 X 100 possible relative eye movements (Yarbus 1967), requiring a map with 10 billion entries: 10 billion synapses comprise only 0.001% of the brain's total.

Acquiring a visual motor calibration map

The utility of a table-based model as a representation for perceptual-motor coordination depends in large part on whether or not it is actually possible to determine the values that constitute the table. Obviously enough, filling such a table in by hand is intractable. To learn the visual motor calibration map, IRV makes camera movements, systematically or at random, and fills in the table with observed correspondences between pixels before and after a movement. For example, if a particular blue pixel appears at location A before movement M, and then at location B after the movement, the system would add an entry for $R(A, B, M)$.

However, the visual motor calibration map is not amenable to such simple learning by observation and memorization, simply because, for any two views before and after a camera movement, the question of which pixel corresponds to which is ill-posed. This is especially true when the pixels are arbitrarily ordered. Regardless of the criteria used to find corresponding points in the images, many spurious correspondences will occur, and true correspondences will often be missed. The color- matching criteria we used, a 5% difference threshold on the red, green, and blue pixel components, generally produced about 100 false correspondences for every true one, and yet missed true correspondences more than half the time.

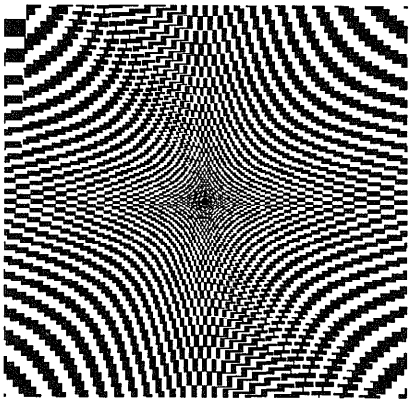

Figure 4: The approximate size and location of the sampling regions in a 400x400 input, alternately colored black and white. Each region is actually square, with some overlap.

Despite these difficulties, a successful method for acquiring the visual-motor calibration map has been developed, described in detail elsewhere (Prokopowicz 1994). Very briefly, our solution to this noisy learning problem considers all apparent correspondences as evidence for a possible true correspondence, and accumulates evidence by repeating the movements enough times. A straightforward application of this idea explodes the size of the table by a factor of N_{pixels}, which is not feasible on IRV's or human hardware. A time-space tradeoff is possible that reduces the growth of the table to only a small constant factor while requiring that the movements be repeated a reasonable number of times.

Although it must cope with massive ambiguity and uncertainty, our algorithm is able to identify truly corresponding image locations after only several hundred examples of each movement, in a space only ten times larger than the map size. For IRV, who makes a movement every 4 seconds, the whole process typically takes two days. For a human, whose eyes move about 10 times more frequently, and using a parallel implementation of the algorithm, the process could be complete in roughly 80 days, even though a human can make about 400 times as many different eye movements.

Off-line calibration in motor coordinates

The visual motor map directly supports tasks that demand perceptual integration, by enabling image prediction and short-term envisioning of old features in the current retinotopic frame. This will be clarified and demonstrated shortly. But the map also makes it possible to interpret the scrambled, distorted geometry of a single retinotopic image, which may contain visible and remembered features. Measuring the geometric relationships in an image is crucial in computer vision whether the camera moves or not.

The interpretation process requires that the system know the mount's pan and tilt angles for each relative movement in the table. These angles can be used to define a motor-based coordinate system for measuring the relative visual angle between any two image locations. If the relation $R(A, B, M)$ holds for image locations A and B, and platform movement M, then the image locations describe lines of sight separated by an angle proportional to M. Using A and B to find the movement M for which they correspond, it is possible to uncover the true two-dimensional geometric relationship between any pair of points, provided that those points correspond to the same line of sight before and after some known camera movement.

We have taken another approach, using only a small subset of all the movements needed for complete coverage, and an off-line calibration process which yields another table that directly supports accurate visual-angle judgments using a motor metric. Instead of consulting the visual motor calibration map to find the motor angle between pairs of image locations, we use the map to assign to each scrambled, non-uniform pixel location a canonical coordinate, in a globally consistent way.

Each entry in the visual motor calibration map constrains the coordinates assigned to a pair of points such that they are separated by an angle equal to the movement angle for which they correspond (fig. 5). The local constraints in the map can not normally be satisfied simultaneously, but the total error can be minimized. We use a numerical simulation of a pseudo-physical process that pair-wise forces points into a configuration satisfying their constraints.

The result is a consistent assignment of a motor-based coordinate to each image location. As the table is filled in and refined with more examples, the constraints more accurately reflect the underlying imaging geometry (fig. 6). Fig. 7 (top) shows a foveal image before and after arbitrary pixel reordering. During map development, the off-line calibration process reorders and redistributes the pixels to their original relative positions. This assignment compensates for any optical distortions, sampling geometry, mechanical alignment (fig. 8), and, in the general case, any arbitrary reordering of the pixels.

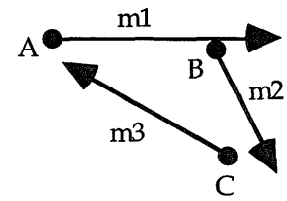

Figure 5: Three mutually corresponding pairs of pixels (A,B) (B,C) and (C,A) whose relative locations are not globally consistent with the movement vectors m1, m2, and m3, over which they have been found to correspond.

Using the visual motor calibration map for perceptual integration across eye movements

The rest of this paper shows how the visual motor calibration map and motor-calibrated image locations can be used for two specific tasks that were designed by psychophysical researchers to measure human capacity for perceptual integration across eye movements: noticing small movements that take place during the saccade itself, and judging large shapes that are presented gradually across a series of fixations.

Stable perception and change detection across saccades A stable world is perceived as stable across saccades under normal conditions, but not when the viewpoint following a movement isn't as expected, nor when something substantially moves during the saccade. These observations motivated the hypothesis that we predict what we will see after a movement and compare it with what we actually get (fig. 9). We tested this feasibility of this hypothesis by using IRV's acquired visual motor calibration map as an image predicting network (fig. 3). During an eye movement, the non-uniform pixels of an image were rearranged, according to the map, to predict the image that would follow the movement (fig. 2). The actual and predicted images were compared pixel by pixel, and those that differed by less than a 20% threshold were considered to match. For a largely stable scene, most of the image is predicted correctly; the scene is perceived as stable. In fig. 9, two small areas are not predicted correctly. These correspond to the area where a mug was displaced slightly on the shelf; IRV noticed this discrepancy and registered its surprise. Humans performed similarly in an experiment that measured the sensitivity to small displacements during a saccade (Bridgeman, Hendry, & Stark 1975). Both humans and IRV

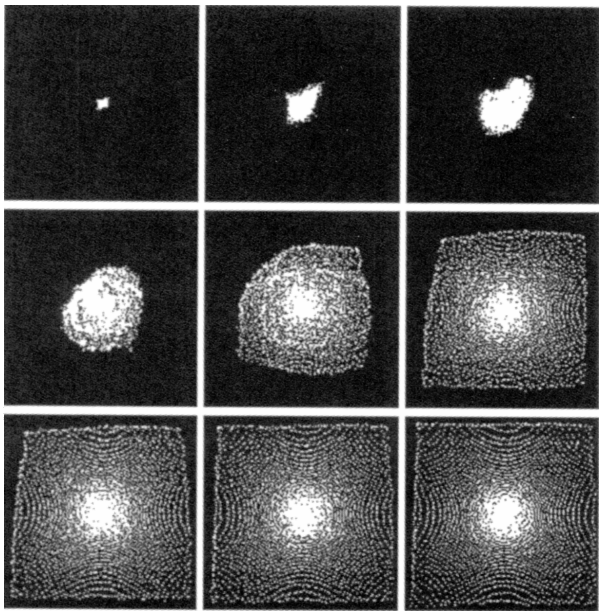

Figure 6: Each pixel of the scrambled, foveated visual representation is assigned a motor coordinate through a pair-wise constraint satisfaction process. Here, the pixels are shown migrating to a pattern that reveals the original sampling distribution (fig. 4). This occurs gradually over approximately 25,000 eye movements, which takes about 24 hours for the robot.

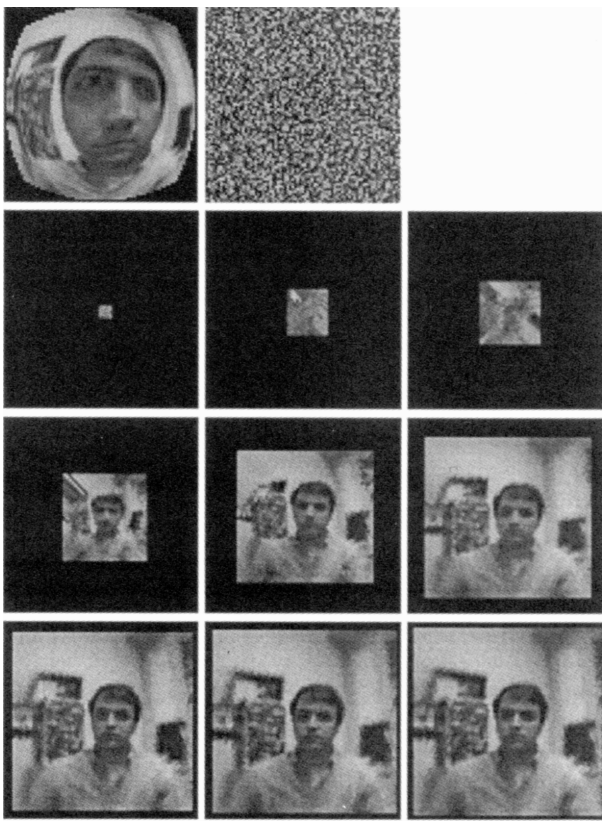

Figure 7: Top row: An example of how IRV's inputs during development are resampled non-uniformly (L) and then reordered (R). Next rows: Each picture shows the same scrambled pixels placed in canonical motor coordinates. As IRV develops an accurate visual-motor calibration map from experience, the assignment of motor-coordinates for distorted and scrambled pixels improves. This sequence represents learning during approximately 25,000 eye movements.

also notice small global scene shifts during a saccade.

Summarizing, through an acquired visual motor calibration map, as embedded in a predictive, retinotopic connectionist architecture, IRV perceives the stability of the external world across camera movements, despite the radically non-uniform changes that result from the movements. Also, IRV can detect small, local scene displacements that occur during eye movements.

Recognition across eye movements Each fixation of a non-uniform sensor gives a highly incomplete view of the world (fig. 1). A crucial component of normal perception under these conditions is recognizing the large-scale relationship among the details acquired from each view. Obviously, this is not the only problem involved in general object recognition, but any theory of recognition that concerns moving, non-uniform sensors must support perceptual integration of large-scale visual geometry across frequent movements.

We propose that when a form is so large that its defining features are not simultaneously visible, a quick series of movements acquires the features foveally, and the features viewed earlier, although no longer recognizable at low resolution, are envisioned in their new retinotopic locations. The basis for visual integration across eye movements is a mental relocation of all visual features into a new frame of reference with each eye movement. This process can apply to any retinotopic feature map. These maps preserve the non-uniform distribution of spatial acuity, because it is not possible to represent all visual information at the highest acuity.

The visual motor calibration map provides exactly the geometric information needed to envision the retinotopic location of an object or feature after an eye movement, with the same spatial resolution of the sensor. We have already seen that simple color features (pixels) can be envisioned in their new retinotopic location. The same mechanism is used here to envision the vertices of a triangle as they are presented on successive fixations. Then, the actual geometry of the triangle is judged using the motor-based coordinates

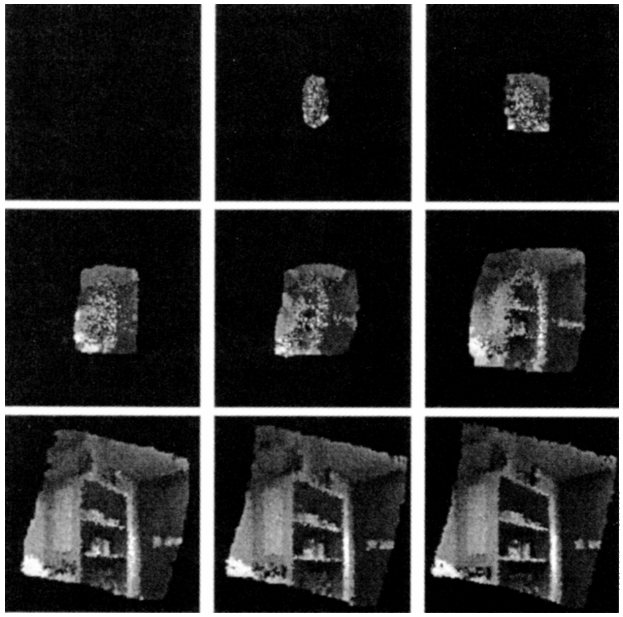

Figure 8: In this sequence, the camera was twisted with respect to the pan and tilt axes. The developing visual motor calibration map gradually uncovers the true imaging and sampling geometry.

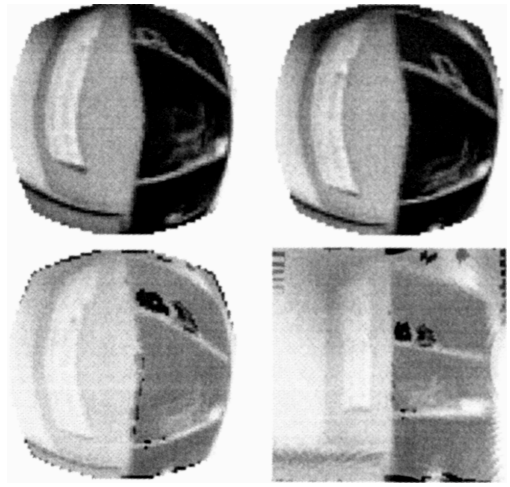

Figure 9: Top: Consecutive, foveal views of a scene. A visual robot will want to know if anything in the scene changed from one view to the next. The comparisons needed to do this are one form of perceptual integration across eye movements. Bottom: During the camera movement, an expected image was predicted; most of the scene was confirmed (faded). The two areas not predicted correctly correspond to where a mug was moved during the camera movement (shown analytically unresampled at right).

of the envisioned vertices, based on the off-line image calibration process described earlier.

The perceptual task is a replication of a psychophysical experiment that measures human accuracy of form integration across eye movements (Hayhoe, Lachter, & Feldman 1990). The task is straightforward: judge if the angle formed by three points of light, presented one at a time on successive fixations, is obtuse or acute. An accurate judgment entails knowing the relative positions of the dots with respect to each other, which in turn depend on the size and direction of the intervening eye movements. Since the presentation is entirely empty except for the single point on each fixation, the subject has no visual cue to judge how much the viewpoint changed; only non-visual, or so-called extra-retinal (Matin 1986), eye position information can be used. The subjects judged the angles to within a threshold of six degrees. If a single stable visual cue persisted across the saccades, accuracy roughly doubled, and was equal to the control case in which all three points were presented simultaneously.

In our experiments, three dots, defining a right angle, were presented one at a time, with small, random camera movement between each presentation. During the camera movement, any previously acquired dots were mentally shifted by IRV into a new, predicted retinotopic position. After the third dot was shown, the angle between the dots was determined using the motor-based coordinates of each envisioned point. In ten trials, the average absolute error in perceived angle was 3 degrees.

Conclusions

We have found that the geometric knowledge required for integrating visual information across camera movements can be represented conveniently as a visual-motor calibration map. The map defines image points that correspond to the same sight lines before and after a particular camera movement. It has an equivalent connectionist network that can predictively shift remembered visual information, during a camera movement, into the new retinotopic reference frame. The map and its network constitute the representational basis for an alternative to theories relying on stable, head-centered reference frames hypothesized to exist in our visual systems. Such eye-position invariant visual responses have yet to found, while predictive activity shifts of the sort proposed here are widespread (Duhamel, Colby, & Goldberg 1992).

The experiments described here show that useful processes demanding perceptual integration can be carried out, in a retinocentric frame, by using learned

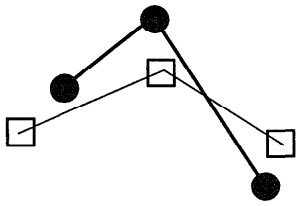

Figure 10: Schematic form-integration problem. Each frame shows the image of one corner of a quadrilateral, relative to the center of view. In order to determine the overall size and shape of the figure, it is necessary to combine information from the four images using knowledge of the intervening eye movements.

visual expectations. Unexpected visual changes that occur during eye movements can be detected by noticing differences between actual and expected pixels; at the same time, confirmed expectations constitute normal stable perception despite frequently shifting inputs. Visual features that are too small and too far apart to resolve simultaneously can be integrated from successive fixations by continually envisioning the retinotopic positions of remembered features, using the same mechanism of learned visual expectations. If the parameters of the eye/camera movements are known, the map can be used to interpret the relative positions of a pair of image points in terms of the movement for which they most closely correspond.

The visual motor calibration map is intrinsically adaptive, since it is acquired from natural, ambiguous visual experience. We have shown how to extrapolate the information in the map to provide a complete calibration of the visual-motor system that accounts for optical parameters and distortions, non-uniform sampling, mechanical misalignments, and arbitrary pixel ordering. In short, the visual motor calibration map can serve as the basis for vision with moving, foveal cameras, providing both a wider field of view and higher spatial resolution.

References

Atkeson, C. G. 1990. Using local models to control movement. In Touretzsky, D. S., ed., *Advances in Neural Information Processing Systems*, 316–323. Morgan Kaufmann.

Bridgeman, B. D.; Hendry, D.; and Stark, L. 1975. Failure to detect displacement of the visual world during saccadic eye movements. *Vision Research* 15.

Duhamel, J. R.; Colby, C. L.; and Goldberg, M. E. 1992. The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* 255(90):90–92.

Faugeras. 1992. What can be seen in three dimensions with an uncalibrated stereo rig. In Sandini, G., ed., *Proceedings of the 2nd European Conference on Computer Vision.* Springer-Verlag.

Feldman, J. A., and Ballard, D. H. 1982. Connectionist models and their properties. *Cognitive Science* 6.

Feldman, J. A. 1985. Four frames suffice: A provisional model of vision and space. *Behavioral and Brain Sciences* 8(2):265–289.

Hayhoe, M.; Lachter, J.; and Feldman, J. 1990. Integration of form across saccadic eye movements. Technical report, University of Rochester.

Matin, L. 1986. Visual localization and eye movements. In Boff, K. R.; Laufman, L.; and Thomas, J. P., eds., *Handbook of Perception and Human Performance*, volume 1. New York: John Wiley and Sons.

Mel, B. W. 1990. *Connectionist robot motion planning: A neurally-inspired approach to visually-guided reaching*, volume 7 of *Perspectives In Artificial Intelligence.* Academic Press.

O'Regan, J. K., and Levy-Schoen, A. 1983. Integrating visual information from successive fixations: Does trans-saccadic fusion exist? *Vision Research* 23(8):765–768.

Pollatsek, A.; Rayner, K.; and Collins, W. 1984. Integrating pictorial information across eye movements. *Journal of Experimental Psychology: General* 113(3):426–442.

Prokopowicz, P. N. 1994. *The Development of Perceptual Integration Across Eye Movements in Visual Robots.* Ph.D. Dissertation, Institute for the Learning Sciences, Northwestern University.

Sparks, D. L., and Porter, J. D. 1983. The spatial localization of saccade targets. ii. activity of superior colliculus neurons preceding compensatory saccades. *Journal of Neurophysiology* 49:64–74.

Stratton, G. M. 1897. Vision without inversion of the retinal image. *Psychological Review* 4:342–360.

van Essen, D. C.; Felleman, D. J.; DeYoe, E. A.; and Knierim, J. J. 1991. Probing the primate visual cortex: Pathways and perspectives. In Valberg, A., and Lee, B., eds., *Pigments to Perception.* Plenum Press, New York. 227–237.

Yarbus, A. L. 1967. *Eye movements and vision.* Plenum Press.