Mr. T Does Motion Parallax

Ben Chandler, Sean Lorenz, Mikhail Panko, and Siddharth Rajaram

Abstract—**Understanding how humans learn to view 3D scenes, gauging depth and distance is an important topic in neuroscience. In this paper we investigate motion parallax by way of Mr. T, a virtually simulated robot with a torso, arms, anthropomorphic neck, head, and cameras for eyes. Using proprioceptive and visual information, Mr. T was able learn how to foveate its camera and pan the head towards a novel end effector position after several training trials. A K-Nearest Neighbors algorithm was used to test Mr. T's learning, however, a theory of cortical networks that are potentially involved in extracting features such as depth and distance in humans is also discussed.**

I. Introduction

THE extraction of depth and distance information from 2D retinal inputs is a difficult and still misunderstood area of neuroscience. The neural computations involved in establishing 3D spatial representations of the outside world rely upon a vast network of visual, parietal, and frontal lobe regions with different regions becoming active over space and time. Humans and other animals use more than retinal inputs to calculate distance and depth, coordinating eye, head, and body movements as well. This framework suggests that in order to process spatial distance from our body to a target, an animal must take into account dynamic variables computed by the brain via multiple gradients of input flow in space and time.

This paper explores a method for calculating distance information implicitly encoded in the arm position to calibrate the 3D information given by head-movement parallax. The robot, Mr. T, as well as the Mr. T Simulator, were used to learn distance calculation along the azimuthal plane in one eye. Using eye, head, and proprioceptive arm position information, Mr. T was able to learn distance and depth cues in space over time.

Although the K-nearest-neighbors (KNN) algorithm used to learn the task is not biologically plausible, several key brain regions (located in the occipital, parietal and frontal cortex) and functional properties of these areas are discussed in order to give a realistic approach to how the brain might be computing the different apparent motion of end effector position at different distances.

Manuscript received May 13, 2008.

II. Simulation Methods

Mr. T is a humanoid robot developed at Boston University's Active Perception Lab which was built to show how computational models could perform behavioral tasks pertaining to vision, in particular. Mr. T accurately replicates retinal image motion during fast macroscopic saccades and during fixational eye movements, and is equipped with two anthropomorphic arms, each with 5 degrees of freedom. Before testing a task on Mr. T, however, simulations were run on the Mr. T Simulator that uses physically-based rendering software and Matlab to replicate in a 3D simulation space the physical movements of Mr. T's eye, head, and arm movements.

First, to reduce the scope of issues associated with parallax, simulations were restricted to the azimuthal (x-y) plane in just one of Mr. T's two eyes (right). Also, to simplify the visual field within the Simulator, a point light source (PLS), acting as the focal point, was affixed to the right index finger of Mr. T's arm. This white dot PLS on the center of the camera's ("eye") field of view is determined by: 1) if the PLS is on the left half of the screen rotate the eye x degrees where x is small enough to avoid making the dot disappear. If this procedure does not move the PLS to the right half of the screen, x is moved again until the PLS is centered in the camera's field of view. Once the PLS is on the right half of the screen, the camera moves left by $x/2$ degrees, $x/4$ degrees, $x/8$ degrees, etc., until it is centered. Note that this entire procedure takes place in eye coordinates; the head and arms stay still during this first task.

Second, the issue of parallax is addressed when the head, first looking straight ahead, pans by a fixed amount, θ degrees to the right, and calculate the amount, ϕ_1, needed to rotate the fovea to the left in order to re-center the camera. Then, the camera is moved θ degrees to the left and a turn to the right, ϕ_2, is calculated. Note that ϕ_1 and ϕ_2 are the *changes* in eye position, thus binocular disparity or vergence information is not being used here. Mr. T's goal is to correlate the changes in eye position with x-y spatial position of the PLS in order to interpolate for un-sampled test locations of the arm in space. Thus, by using eye, head, and arm position information for a certain fixed arm position, Mr. T should be able to learn how to foveate its camera and pan the head towards a novel end effector position after

numerous training trials.

We used KNN classification (with k = 1) trained on 15 sample points. Each training point was obtained by positioning the end effector in a new position by varying the elbow and wrist joint angles. The locations of the PLS are plotted below in Figure 1. Our KNN classifier works by selecting the cluster with the smallest difference between the trained and computed eye-pan angles.

A natural extension to this model would be to find the two closest clusters to a given point and interpolating linearly, with weights proportional to the differences of the computed and trained eye-pan angles. In addition we expect that with more training points, the performance will improve significantly and could easily be adapted to learn 3-dimensional depth cues.

Furthermore, it would be worth exploring the representation of depth in visual and parietal cortex to formulate a better interpretation of these abstract clusters. Either a topographic layout of depth representation or information about the grain of sampling could be added on to the model to explore the effects of such an organization and to lend additional biological plausibility.

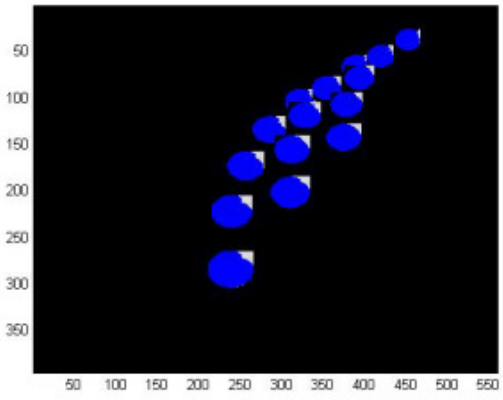

Figure 1: Locations of the PLS as seen by Mr. T's right eye during training of the KNN classifier

III. Simulation Results

The shoulder azimuth was kept constant at 90°, the shoulder elevation was kept at –45° and neck azimuth was also held constant at –45°. These values were chosen in order to get the hand as close as possible to Mr. T's head, as parallax cues are strongest when the object is close to the eyes – this is evident mathematically and observed in psychophysics. The closest hand position had a wrist angle of 40° and an elbow angle of 100°, as depicted in Figure 2.

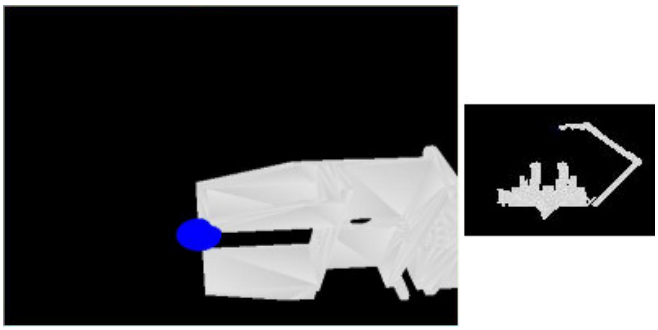

Figure 2: (left) View from Mr. T's right eye looking straight at the closest hand position, (right) Overhead view of Mr. T. at the same position

The furthest hand position had a wrist angle of 0° and an elbow angle of 90°, as depicted in Figure 3.

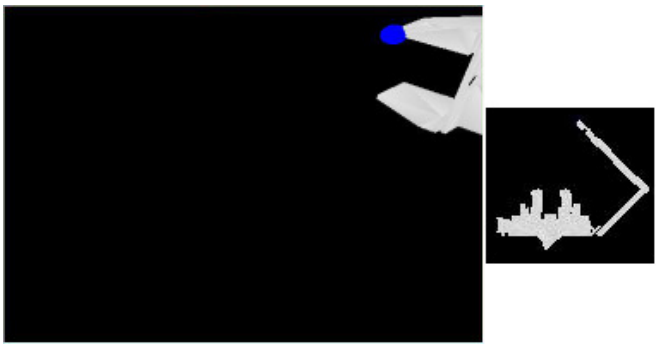

Figure 3: (left) View from Mr. T's right eye looking straight at the furthest hand position, (right) Overhead view of Mr. T. at the same position

We ran the simulations with our fixation algorithm and computed the relative angular eye displacements after neck pans. After training, we ran the algorithm on test points and recorded the eye-pan angles. The eye-pan angles for one of the examples for a set of three fixations are shown below in Figure 3. The elbow angle was 97°, and the wrist angle 21°.

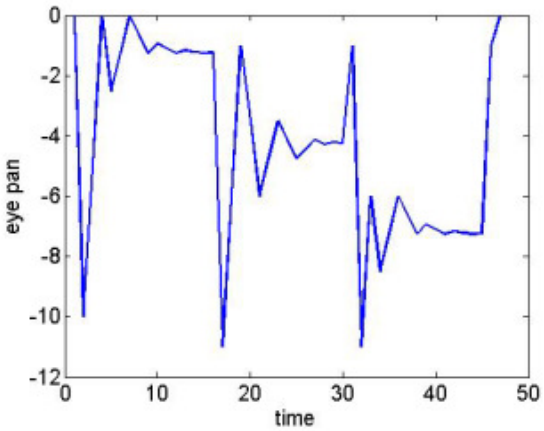

Figure 4: Eye-pan angle values during a set of three fixational eye movements

Note the three large dips in the above graph – these correspond to the first large eye movements, followed by smaller jittering (by design of the algorithm) leading to foveation (i.e. centering of the PLS on the image) at an eye-pan angle of approximately -1°, as shown in Fig. 4. Then around time = 15, we pan Mr. T's neck 3° to the right and fixate on the PLS once again, leading to an eye-pan angle of approximately -4°. Finally, the neck is panned another 3° to the right and this process is repeated to yield and eye-pan angle of about -7°.

These two values of eye-pan angles (approximately -4° and -7°) are then used as the input to the trained KNN classifier. The classifier outputs an elbow angle of 95° and wrist angle of 20°, which is a very close approximation to the actual values of 97° and 21° respectively.

IV. Discussion

A. Mr. T's Learning

The most prominent feature of our algorithm was to enable Mr. T to perform visuo-motor transforms, by learning the mappings of eye-pan angles directly to joint angles required for performing reaches and manipulating objects. Note that at least in the peripersonal space, visuo-motor transformations of this kind are believed to take place in parietal cortex, as will be discussed in the next section.

Evaluated independently of the ARPI modeling package, the fixation, parallax, and classification algorithms performed well. When combined, numerous ARPI bugs crippled our simulation effort.

Temporal issues when interacting with the full ARPI simulation suite caused a first round of issues. The ARPI functions update Mr. T's state asynchronously, a behavior which is not well-documented or well-handled. Since the API does not provide callback or synchronization functionality, the only way to ensure that a robot state update had completed after a command was to force a hand-coded delay. Without such a measure, queries to the API issued during a state update would return inconsistent results. Even with a hard-coded delay, behavior was still stochastic. Under more CPU-intensive conditions, the hand-coded delay was not sufficient to ensure synchronization.

Besides undocumented temporal behavior, numerous inconsistencies and issues with the base API functions caused significant problems. When building the fixation algorithm, we implemented both vertical and horizontal eye motion. However, the API call to retrieve the current angle of the eye returned a consistent garbage value. Without valid feedback, the vertical eye fixation code looped infinitely.

Taken alone, none of the bugs were critical. The combination, however, significantly dampened our modeling efforts. The model we produced had some learning capability, but was riddled with workarounds designed to handle the numerous ARPI problems.

B. Cortical Networks of Depth and Distance

Results for the simulation show that Mr. T can adapt its behavior in order to perform the simple task discussed here, however, if Mr. T were to perform more biologically-realistic computations one must look into the networks of the brain for clues. Mr. T is an embodied robot which receives "proprioceptive feedback" from its eye, head, and arm positions as well as visual input from the cameras themselves, thus there are several cortical connections of interest. Below is a simplified overview of several areas which must be accounted for when considering a future, biologically plausible adaptive model for Mr. T.

First, retinal input is received in 2D and relayed through the LGN where it continues to process visual information in V1. Information from V1 then ascends into V2 and higher visual areas such as V3, V4, V6/V6a, MT, and MST. Many of these regions are reciprocally connected. In fact, it is believed a corollary discharge of the eye movement command updates an internal representation when the eye moves [1]. From these (and other) areas, features such as color, orientation, edge detection, figure-ground separation, motion, depth, distance, and a host of other difficult visual computation occurs via the supposed dorsal and ventral streams of the occipital cortex. For this paper we are more concerned with the dorsal "where" or "how" stream since we are interested in the establishment of 3D spatial representations needed for properly calculating distance from the eye/head to an end effector target position.

One issue that must be addressed as well is the issue of varying representations of space in the brain. Visual cortex codes with retinotopic organization, somatosensory cortex is somatotopically organized, posterior parietal cortex is said to represent space as interactions between many modalities, and motor, premotor, and supplementary motor areas apparently have both spatial and motor representations. Thus it seems that association areas are ideally suited for spatial representation due to the ability of combining information from many modalities. Burnod et al. [2] have constructed an integrated framework for parieto-frontal coding of reaching that combines neural inputs based on four types of reach-related signals: retinal, gaze, arm position/movement direction, and muscle output). It is believed that this sort of

sensory-motor information flow based on combinatorial domains in the cortex can also be an important way of viewing the extraction of depth and distance from 2D images.

For Mr. T's purposes, there is information from the end effectors which act as motor-related limb feedback – as does information pertaining to eye and head position. The motor-related vectors are received in primary somatosensory cortex and must be sent to other areas in order to combine its information with that entering from the actual visual input to V1 and other upstream occipital region processing. The area most often sited as being concerned with an allocentric, 3D spatial representation of external space is the posterior parietal cortex (PPC). It may be that a "spatiotopic" representation is achieved by way of multiple sensory modality inputs [3]. This process is said to occur in PPC's interactions with primary modality regions and reciprocal frontal cortex connectivity.

Some of the key PPC, more specifically the intraparietal sulcus (IPS) regions shown in Fig. 5, involved in establishment of 3D spatial representation will now be discussed. First there is the parietal occipital area (PO), also known as V6/V6a, which is not shown but is located just above the medial intraparietal (MIP) region. V6 receives inputs from V2, V3, V3a, and MST whereas V6a receives inputs no V2 input, weak V3/V3a inputs, and stronger V5 inputs. It also sends strong projections to MIP and seems to code reaching activity retinotopically.

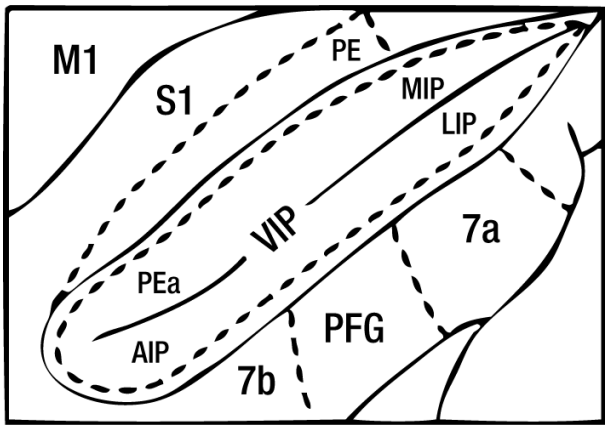

Fig. 5. Detail of the intraparietal sulcus where AIP is more anterior towards frontal cortex and LIP is more posterior towards occipital cortex. Note: dashed line represents opening of sulcus for better view.

Also not shown in Fig. 5 is the caudal intraparietal (CIP) area which is located just posterior and dorsal to lateral intraparietal (LIP) area. CIP receives inputs from V3, V3a and V4 in particular. It is said to be involved in 3D analysis of object features, containing surface-orientation-selective as well as axis-orientation-selective neurons.

LIP receives inputs from PO, ventral intraparietal (VIP) area, MT, MST, V2, V3, V4, supplementary motor area (SMA), and frontal eye fields (FEF). LIP is said to be an extremely important area, as shown by its numerous inputs, with much research showing its significance for saccade planning.

MIP receives inputs from PO, VIP, and dorsal premotor cortex (PMd). Just as LIP is known to be vital for saccade planning, MIP is vital for arm reaching movement planning.

VIP receives inputs from MT and MST, as well as numerous primary somatosensory areas. It is known to be highly polymodal and sensitive to direction of sensory movement (be it tactile or visual) in particular. It has also been hypothesized that the VIP-F4 circuit is possibly involved in peripersonal space coding for movement [4].

Finally, anterior intraparietal (AIP) area receives inputs from CIP and the hand-related motor areas of F5. It is known to be involved in coding of 3D object characteristics primarily for grasping-related visuomotor transformations.

So how does all this tie together? It is believed that each parieto-premotor circuit is dedicated to a specific sensorimotor transformation [5]. For example, for the AIP-F5 circuit, a multiple pragmatic description of 3D objects is provided by AIP, proposing several grasp possibilities to F5. F5, then, selects the most appropriate grasp based on prior context and rule-association and sends it to F1 for motor output as well as back to AIP neurons coding that selected grip in order to keep them active during movement execution. This same sort of circuitry occurs in the MIP-PMd circuit for reach planning and the LIP-FEF/SMA circuit for saccade planning.

One more issue needs to be addressed concerning establishment of 3D spatial representation – the relation of time and space in visual perception. The inner representation of space cannot be considered solely via static investigation since animals are *actively* viewing and acting in their environment. These inner representations are also dependent upon dynamic variables computed by the brain on the basis of "gradients of input flow" over space and time [6].

The views of PPC and visual cortex above suggest that visuomotor transformation relies on parallel collaboration between frontal, parietal and occipital areas which form unique parietofrontal (and parieto-occipital) circuits where sensory and motor signals are integrated. This idea is also consistent with the high level of plasticity in PPC due to its need for adaptation to anterior and posterior inputs needed to create 3D representations of space over time. Such a view also sees space coding as perhaps being a secondary result of

the activity of these circuits where spatial location of an object is coded according to the animal's motor planning purposes. Thus, multiple space representations are constructed that, if true, should be seen experimentally in lesions of that specific circuit. Were such a theory to be true this suggests that existence of a single, multipurpose space are is inconsistent with anatomical and neurophysiological data.

Using this hypothesis as a framework, then, we can propose a method for how Mr. T would perform the simulation task mentioned early in the paper within a biologically-plausible adaptive model. First, saccadic movement to the PLS on the tip of the end effector may primarily occur in the LIP-FEF circuit which also involves the SMA and superior colliculus (SC). Then, the head must move to where the fovea/camera was previously centered on the PLS; proprioceptive information from Mr. T's neck rotation must be used to move an appropriate degree amount either left or right. In the brain, somatosensory neck inputs may be sent to VIP which is direction-specific across modalities, which then requires the eye to re-adjust in order to center the PLS in its foveated vision space. This requires the LIP area to adjust, learning a new saccadic movement in conjunction with the head movement. This task also requires proprioceptive information from the relative position of the arm in space. Putting it all together, arm feedback is also sent to the PPC for analysis in primarily the MIP-PMd circuit which is also linked to the V6/V6a areas for updating of future arm movements in space. Thus, Mr. T can adaptively learn novel saccades and head movements to a PLS attached to an arm over time based on visual input from the eye and proprioceptive feedback from eye, head, and arm position.

ACKNOWLEDGMENT

Thanks to Mr. T for your continued encouragement and willingness to learn. We salute you and your legless form.

REFERENCES

[1] Merriam, E. P. and Colby, C. L. (2005). Active vision in parietal and extrastriate cortex. *Neuroscientist*, 11(5):484-493.

[2] Burnod, Y., Baraduc, P., Battaglia-Mayer, A., Guigon, E., Koechlin, E., Ferraina, S., Lacquaniti, F., and Caminiti, R. (1999). Parieto-frontal coding of reaching: an integrated framework. *Exp Brain Res*, 129(3):325-346.

[3] Gardner, J. L., Merriam, E. P., Movshon, J. A., and Heeger, D. J. (2008). Maps of visual space in human occipital cortex are retinotopic, not spatiotopic. *The Journal of Neuroscience*, 28(15):3988-3999.

[4] Luppino, G., Murata, A., Govoni, P., and Matelli, M. (1999). Largely segregated parietofrontal connections linking rostral intraparietal cortex (areas aip and vip) and the ventral premotor cortex (areas f5 and f4). *Experimental Brain Research*, 128(1):181-187.

[5] Matelli, M. and Luppino, G. (2001). Parietofrontal circuits for action and space perception in the macaque monkey. *NeuroImage*, 14(1):S27-S32.

[6] Frégnac, Y., René, A., Durand, J. B., and Trotter, Y. (2004). Brain encoding and representation of 3d-space using different senses, in different species. *Journal of Physiology-Paris*, 98(1-3):1-18.