Autonomous and Interactive Improvement of Binocular Visual Depth Estimation through Sensorimotor Interaction

Timothy A. Mann, *Member, IEEE,* Yunjung Park, Sungmoon Jeong, Minho Lee, and Yoonsuck Choe, *Member, IEEE*

***Abstract*—We investigate how a humanoid robot with a randomly initialized binocular vision system can learn to improve judgments about egocentric distances using action and interaction that might be available to a human infant. First, we show how distance estimation can be improved autonomously. We find that actions that maintain invariant distance are a powerful tool for exposing estimation errors. These errors can be used to train a distance estimator. Furthermore, the simple action used (i.e. neck rotation) does not require high level cognitive processing or fine motor skill. Secondly, we investigate how interaction with humans can further improve visual distance estimates. We find that human interaction can also improve distance estimates for far targets compared to autonomous learning without human interaction. Together these experiments suggest that both action and interaction are important tools for improving perception.**

***Index Terms*—vision, depth estimation, autonomy, learning, action, perception.**

I. Introduction

HOW can humans or animals learn to make sense of the data collected by their sensory organs? There is a lot of noisy, messy data and very little obvious meaningful information. For example, the distance estimation problem requires an embodied agent to estimate the distance from the agent's body to a target object. How does the agent learn to make sense of these sensory signals to predict the distance to the target object (see figure 1)?

Our hypothesis is that humans and animals employ several strategies for detecting inconsistencies in its depth estimates and integrates this information together. First, action can improve distance perception by exploiting perceptual or physical invariants. Second, interaction with other social agents can improve distance perception by providing strong cues to the learning system that might be difficult to acquire autonomously.

To investigate our hypothesis, we experiment with the egocentric distance estimation problem on a humanoid robotic platform with a binocular vision system. The objective of the robotic system is to make accurate estimates of the depth from between its two cameras to a target object. Our objective is to learn how an autonomous embodied system can learn to make enough sense of its visual system so that it can make distance estimates to target objects that are accurate enough to facilitate interaction.

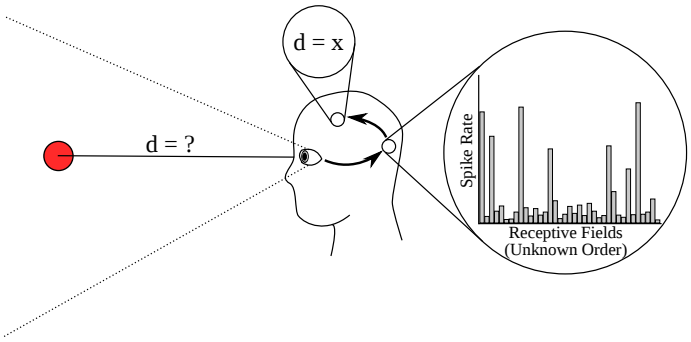


Fig. 1. To estimate distance, sensory stimulus from the retina is encoded by visual receptive fields. The brain must then use these neural responses to predict a distance estimate. How does the brain learn a concept of distance from seemingly arbitrary neural spike patterns?

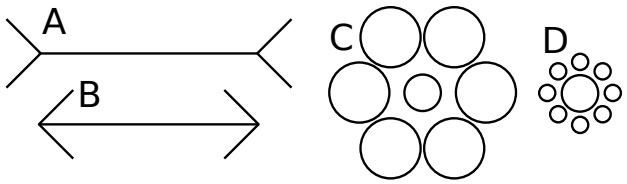


Fig. 2. A and B demonstrate the Müller-Lyer illusion. Although the center line of A and B are the same length, the direction of the arrows on the ends of A makes the top line appear longer, while facing the arrows in the opposite direction (B) makes the bottom line appear shorter. C and D demonstrate the Titchener circle illusion. Again, although the circles in the center of C and D have the same diameter, C appears smaller than D due to the circles surrounding their periphery.

Several psychological studies have discovered evidence that action is unaffected by certain visual illusions or the effect of the illusion is measurably reduced in motor responses. These results support the dual visual pathways hypothesis where the consciously accessible visual pathway is affected by the visual illusion and the visual pathway tightly coupled with motor control is unaffected. For example, the grasp aperture of human subjects is not found when presented with the Müller-Lyer illusion (depicted in figure 2 A and B) [1]. The Titchener

T. Mann and Y. Choe are with the Department Computer Science and Engineering, Texas A&M University, College Station, TX, 77843 USA e-mail: {mann23, choe}@tamu.edu

S. Jeong and M. Lee are with the School of Electronics Engineering, Kyungpook National University, 1370 Sankyuk-Dong, Puk-Gu, Taegu 702-701 Korea e-mail: jeongsm@ee.knu.ac.kr, mholee@knu.ac.kr .

Y. Park and M. Lee are with the School of Robotics Engineering, Kyungpook National University, 1370 Sankyuk-Dong, Puk-Gu, Taegu 702-701 Korea e-mail: yj-park@ee.knu.ac.kr,mholee@knu.ac.kr .

Manuscript received January 15, 2012;

circles illusion (depicted in figure 2 C and D) has also not found in grasp aperture when human subjects move their hand to pick up the center circle [2]. The Müller-Lyer illusion can be considered a 1-dimensional visual illusion because it influences our perception of visual lines. The Titchener circles illusion can be considered a 2-dimensional illusion because it influences our perception of size. By analogy, we can consider distance estimation as a 3-dimensional perceptual problem with many opportunities for illusion to obscure the true distance. Can action also minimize the perceptual error in distance perception?

There is strong experimental evidence that action is critical in the development of depth perception. For example, in a classic experiment, Held and Hein [3] placed two neonatal kittens on a circular harness surrounded by uniform visual stimulus. An active perceiving kitten caused both kittens to move, while the passively observing kitten was exposed to identical stimulus. In subsequent tests of visually guided movements, the active perceiving kitten performed normally, while the passive observer had deficiencies. This result suggests that self-driven action is important in the development visual perception. For a discussion on recent work related to the importance of action to depth perception, the reader is referred to [4].

There is also a wealth of literature supporting the hypothesis that humans understand the actions of other humans by simulating those actions. Functional MRI studies have discovered evidence that regions of the brain implicated during motor control are active during perception of individuals engaging in fine motor skills (e.g. [5, 6]). Furthermore, the discovery of mirror neurons in both humans and monkeys [7] further implicates a tight coupling between perception and action.

In this paper, we investigate the importance of the action-perception cycle in the problem of visual distance estimation. Although there are many heuristic strategies for visual distance estimation, we focus on binocular distance estimation because it has the potential to accurately estimate distance under a wide range of circumstances.

Binocular distance estimation can be broadly divided into two problems: the correspondence problem and distance estimation from information extracted from the images. Much of the psychological and computer vision literature has focused on the correspondence problem of binocular vision systems [8, 9]. The correspondence problem is the the problem of identifying pixels or regions from both images that correspond to the same physical entity. We focus on the distance from the extracted information. The main problem is that the extracted information may be biased. We develop a learning system that can correct for this bias.

To investigate the role of action and interaction in depth estimation we perform two sets of experiments with a binocular robot platform. The first set of experiments consider how an embodied agent can autonomously learn to estimate distance to a target object. The second set of experiments investigates how human interaction can further improve our autonomously learned distance estimator.

The main contribution of our work is support for the hypothesis that action and interaction improve perception. Specifically, we show how an autonomous, embodied agent can use perceptual and physical invariances to expose inconsistencies in its distance estimates that can be used to train a distance estimator. We also show that interaction with humans can further improve distance estimation by providing information that would otherwise be difficult to obtain autonomously.

The rest of this paper is organized as follows. In section 2, we provide a detailed description of the problem faced by our autonomous agent. In section 3, we explain our approach and experimental setup. Section 5, presents our experiment and results on autonomously learning to estimate distance. Section 6, describes our human interaction experiment. Section 7, discusses the importance of our findings and suggests future work. Section 8, summaries and concludes.

II. The Problem

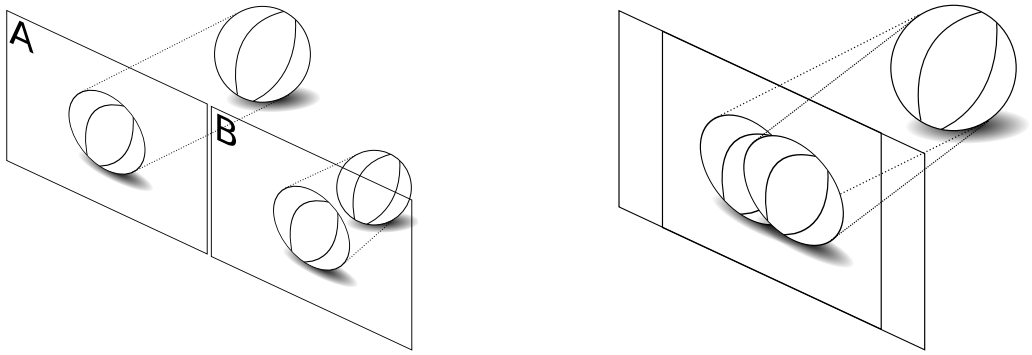
Distance estimation is a nontrivial problem. There are many heuristics for distance estimation that work well under special assumptions but fail under others. The main problem with estimating distance from visual images is that different physical settings can give rise to identical visual images (see figure 3a). Because of this, distance estimation is an ill-posed problem. This is similar to the problem of trying to solve a singular set of equations. There simply is not enough information to identify a unique solution. For this reason, we investigate learning systems with binocular vision. Because a binocular vision system can make two simultaneous observations of the same target object from different perspectives, this information can be used to triangulate the distance from ~~the~~ observer to ~~the~~ target (see figure 3b).

It is worth noting that using binocular vision for distance estimation is related to the use of motion parallax. The main idea behind using motion parallax for distance estimation is to acquire images of a scene from multiple perspectives. Indeed, there is evidence that humans treat these two techniques similarly [10]. By studying binocular distance estimation we may learn valuable techniques for applying motion parallax on autonomous systems. However, we focus solely on binocular distance estimation in this paper.

A. Correspondence Problem

Binocular systems produce two simultaneous images of a target object. This has the advantage of providing the agent with more data, but it comes at a cost. The problem is that regions of both images need to be matched to each other. However, it is not clear exactly how this should be done because the images from the left and right cameras may differ considerably due to to changes in illumination and perspective.

This problem however, has been studied aggressively in the fields of psychology and computer vision (e.g. [8, 9, 11, 12, 13]). For the purposes of this study we will use a biologically inspired attention model presented in [12]. Adopting this mechanism allows us to focus on the other problems faced by autonomous distance estimating systems.



(a) (b)

Fig. 3. (a) Two different physical settings can give rise to identical visual images. (b) This problem can be partially resolved by comparing two or more images of the same physical setting taken from different locations. If the disparity between the two cameras is known, triangulation can be used to estimate the target object's distance.

B. Autonomous Units

Another problem is faced by our autonomous agent. What units of measurement should an autonomous binocular system use? Certainly infants and animals have no notion of the standard units of distance, such as inches and centimeters, that we take for granted. Instead an autonomous learner with a binocular vision system must develop its own unit of measure. But what should the agent base its decision on? Later, we argue that the body itself can provide useful units of measure.

C. Learning to Estimate Distance

Finally, the main problem that we focus on is how to learn to estimate distance accurately. There are three parts to this problem.

1) As an embodied agent, how should that agent interact with its environment to gain information needed to learn to estimate distance?
2) Once information has been collected by the agent, how should the agent use that information to derive accurate distance estimates?
3) How can the agent evaluate whether or not its distance estimates are accurate?

III. Approach

In this section, we explain our approach to the problems described above and detail our experimental setup.

A. Robot Platform & Experimental Setting

We modified an Aldebaran Nao humanoid robot by mounting left and right cameras on its head (figure 4). The robot provides a convenient platform for interaction, while the two cameras allow us to capture images of the environment.

In each experiment a target object is placed in the intersection of the visual fields of the left and right cameras (figure 5). The agent must detect the target in the images captured from the left and right cameras and use information about its location in the images (i.e. pixel coordinates) to estimate the distance of the target object from the robot's head. We also allow the robot to interact with the environment either by rotating its neck or moving its arm. Later we will show how this limited ability to act is critical for learning to estimate distance accurately.

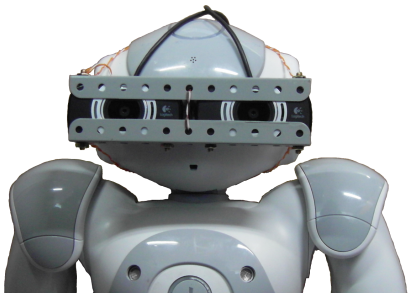


Fig. 4. Aldebaran Nao humanoid robot with mounted cameras.

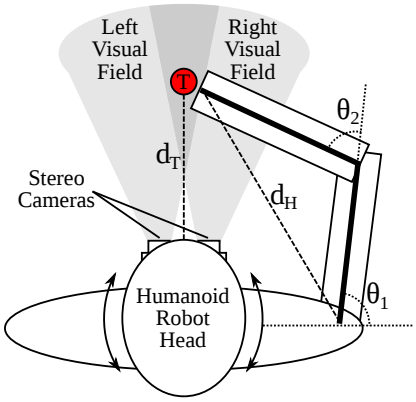


Fig. 5. A target T is placed in the intersecting visual field of the left and right cameras of a humanoid robot. The robot can also interact with the scene by moving its arm or rotating is neck.

B. Attention Mechanism & Correspondence Problem

To detect the location of the target object from the left and right images, we use the biologically inspired attention mechanism described in [12].

When the human eye searches a natural scene, the left and right eyes converge on an interesting area by action of

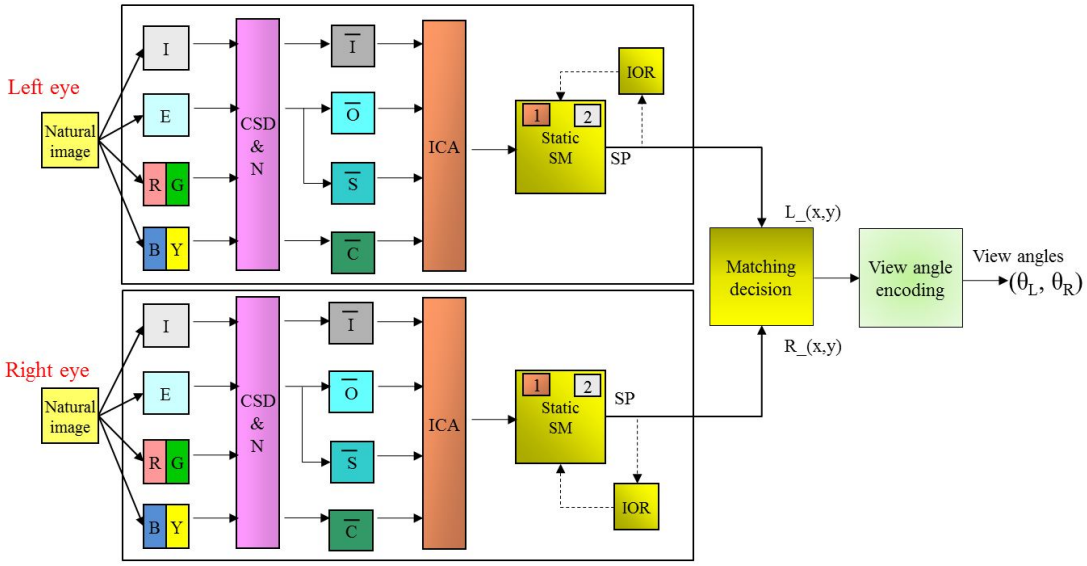


Fig. 6. Bottom-up and top-down visual attention saliency map model. I: intensity image, E:edge image, RG: red-green opponent coding image, BY: blue-yellow opponent coding image, $CSD\&N$: center-surround difference and normalization, $\bar{I}$: intensity feature map, $\bar{O}$: orientation feature map, $\bar{S}$: symmetry feature map, $\bar{C}$: color feature map, ICA: independent component analysis, $StaticSM$: static saliency map, SP: saliency point, 1: first saliency point, 2: second saliency point, $L\_(x, y)$: salient location (x, y) in left image, $R\_(x, y)$: salient location (x, y) in right image, IOR: inhibition of return, angle to the target extracted from the left (θ_L) and right (θ_R) camera images.

the brain and the eyes [11, 12, 13]. With bottom-up (or image-based) processing, the human visual system determines salient locations obtained from features that are based on the statistical information of an input image. Based on Treismans feature integration theory [14], Itti et al. [15] and Park et al. [16] used primitive features such as intensity, orientation, symmetry and color information to construct a bottom-up SM model as shown in Fig. 6. The four features are constructed by a Gaussian pyramid with different scales of Gaussian kernel functions, which mimic the on-center and off-surround roles of the lateral geniculate nucleus (LGN). The center surround difference and normalization (CSD & N) mimics the on-center and off-surround processing and redundancy reduction is performed by independent component analysis (ICA). The LGN and primary visual cortex is implemented in the model as the difference between fine and coarse scales in Gaussian filtering. Various features from the CSD & N processing are useful to detect saliency parts in complex real world images. We regarded a localized area with the highest intensity values in each bottom-up SM as the most salient regions to be analyzed for attending the same location by both cameras. After calculating a suitable size for candidates areas based on entropy maximization [12, 17], to prevent it from being a repetitively attended region in the vision system, the localized region is masked by an inhibition of return (IOR) function [16]. Then the vision system continuously searches for a new localized region by the above procedure. To solve the correspondence problem between the two saliency maps, the camera with higher saliency region is selected as the master and salient regions of the slave image are compared to identify the corresponding region by considering the single eye alignment hypothesis [11, 13]. Comparing the most salient values within selective attention regions in two camera images, we can adaptively select the master eye that has a region with the most salient value. Then, we can obtain an estimate of the radians of two angles (θ_L, θ_R) of the center of the salient location (x, y) in two cameras by simple triangular equation with focal length and CCD width of two camera [11]. Those two angle estimates $(\tilde{\theta_L}, \tilde{\theta_R})$ are used to estimate the depth information of the salient location, where the tilde above the thetas indicates that these are estimates of the true angles.

C. Autonomous Units & Consistent Distance Estimators

What are the properties of a good distance estimator? Since embodied agents are working with noisy sensory data we cannot expect the agent to always predict the true distance. At best we can hope to obtain a distance estimator that is correct on average. Therefore we are interested in distance estimators whose expected value is close to the true distance.

Given $\tilde{\theta}_L$ and $\tilde{\theta}_R$, we define a consistent distance estimator $\hat{d}(\tilde{\theta}_L, \tilde{\theta}_R)$ to have the following property

$$E\left[\hat{d}(\tilde{\theta}_L, \tilde{\theta}_R)\right] = \beta_1 D + \beta_0 \quad (1)$$

where $\hat{d}(\tilde{\theta}_L, \tilde{\theta}_R)$ is the distance estimate, $E[\cdot]$ is the expected value operator, $\beta_1 \in \mathbb{R}^+$ is a positive scalar and $\beta_0 \in \mathbb{R}$ is any real number, and D is the distance in our preferred unit of measure. For clarity of exposition, we will use centimeters throughout this paper.

This objective requires the expected value of a distance estimator to be a linear transformation of distance measured in a standard unit of measure. Requiring $\beta_1 > 0$ ensures that the distance estimator increases with distances and avoids the awkward possibility of $\beta_1 = 0$, so that distance estimators that always output zero are considered inconsistent. But this

objective raises the question, how can an autonomous agent evaluate for itself, whether or not its distance estimation process satisfies this objective? Later, we show that actions that maintain the distance from the robot to the target can be used to learn whether or not the robot's distance estimation process is inconsistent.

D. Learning Framework

We would like an autonomous, embodied agent to learn a general rule for estimating distance to a target object from features extracted from two disparate camera images. These features are potentially noise and are likely to have a nonlinear relationship with the quantity that the agent is trying to estimate (i.e. distance). To maintain any hope of extrapolating (or even generalizing) results of a learned distance estimator to targets at distances that were not in the training set, our learning agent must consider a restricted family of functions.

We assume that there exists an unknown, stochastic function of distance $\mathcal{X} : \mathbb{R}^+ \rightarrow X$, and the learning system is given a parametric model $f : X \times \Psi \rightarrow \mathbb{R}$ for distance estimation, where X is the set of possible features extracted from the binocular images of the target and Ψ is the parameter set. We also assume that there exists $\psi^* \in \Psi$ such that for any distance D in centimeters

$$D = \beta_1 E\left[f(\mathcal{X}(D); \psi^*)\right] + \beta_0 \tag{2}$$

for some $\beta_1, \beta_0 \in \mathbb{R}$. In other words, the expected value of the estimated distance given the optimal argument ψ^* is a linear transformation from the true distance in centimeters. This satisfies our consistency objective (Eq. 1).

The learning objective is to find $\hat{\psi} \in \Psi$ such that for all $D \in \mathbb{R}^+$

$$\hat{\psi} \approx \arg\min_{\psi \in \Psi} E_{x \sim \Pr[\mathcal{X}(D)]}\left[\left(f(x; \psi^*) - f(x; \psi)\right)^2\right] \tag{3}$$

that approximately minimizes the sum squared error. However, because we only have access to a finite number of samples, the learning system needs to estimate Eq. 3 using the following

$$\hat{\psi} \approx \arg\min_{\psi \in \Psi} \sum_{i=1}^{m} \left(d_i - f(x_i; \psi)\right)^2 \tag{4}$$

where $D_i \in \mathbb{R}^+$ for $i = \{1, 2, \ldots, m\}$ are true distances, $x_i \sim \mathcal{X}(D_i)$ and $d_i = \frac{D_i}{\beta_1} - \beta_0$. In the next section, we address how an autonomous agent can gain access to "true" distances by inventing its own unit of measurement.

IV. Autonomous Improvement of Depth Estimation

In this section, we discuss two approaches for learning an accurate depth estimator. Our objective here is to determine when and how an autonomous system can learn to accurately predict the distance to target objects. In the first subsection, we discuss learning to maintain perceptual invariance. In the second subsection, we explain how the invariance of physical size can be used to train a binocular distance estimation system.

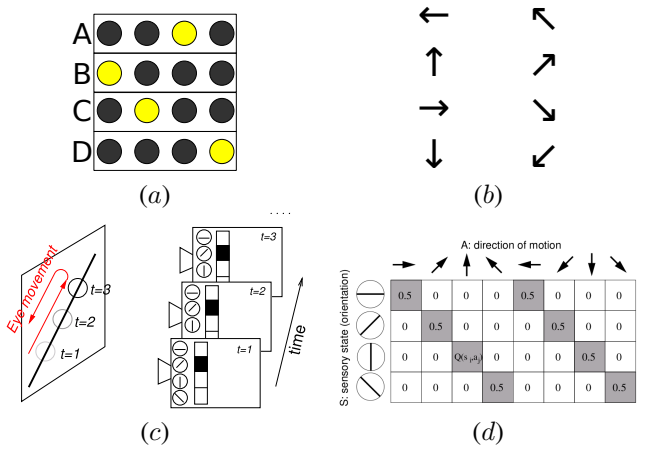


Fig. 7. (*a*) Seemingly arbitrary patterns of lights. What is their intended meaning? (*b*) A set of actions. What are their sensory consequences? (*c*) The environment from perspective of the external observer. If the agent moves diagonally along the line it can maintain the same sensory signal. Adapted from [18]. (*d*) A table expressing the meaning of sensory states through actions that transition to the same sensory state. Adapted from [18].

A. Sensory Invariance Driven Action

A key inspiration for our approach is the idea of Sensory Invariance Driven Action (SIDA). Choe et al. [18] introduce the concept of SIDA to explain how the brain can learn the meaning of sensory stimuli. Because the brain does not have direct access to external stimuli, a critical problem faced by the brain is to understand the meaning of complex neural spiking patterns and to use those patterns to make decisions about how to act in the world. This problem is similar to being shown seemingly arbitrary patterns of light (figure 7a) and being asked to explain what each pattern represents. This problem seems impossible, but it turns out that we can ground the meaning of each pattern with action [18]. Suppose the agent is also presented with a set of actions as in (figure 7b). By experimenting with these actions, the agent can learn the sensory consequences that pertain to the external environment, which in this case is a camera moving over an image (figure 7c). The actions in this case correspond to movement in different directions. The critical insight of Choe et al. [18] is that learning to act in a way that maintains sensory invariance is a useful mechanism for learning the meaning of sensory stimuli. In the example of figure 7, by setting the learning systems objective to maximizing invariance in the sensory stimuli, a mapping from sensory states to actions is learned (figure 7d). This mapping describes the sensory states in terms of actions and the actions in terms of sensory states.

In the next section, we learn to maintain perceptual invariance in situations where, although the robot has physically moved, its action should, in principle, not alter the true distance to the target. If the robot's perceived distance estimate differs, then we can use this difference as an error signal for training.

B. Distance Invariance

One way to improve autonomous distance estimates, is to acquire several training samples that, in principle, should have

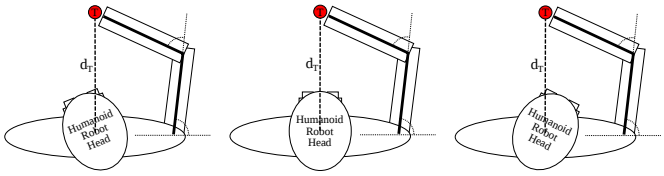


Fig. 8. If the origin is the axis of neck rotation, then rotating the neck does not change the distance to the target, even though the distance from the left and right cameras may have changed.

identical distance from the agent. As long as every sample has identical distance from the observer, the agent can invent a distance unit and apply that to all of the samples. Using these samples would allow us to eliminate parameter choices $\theta \in \Theta$ that produce inconsistent distance estimators. Now if we have a strong enough parametric model f, we can generalize to other distances. The main question is: How can the agent acquire samples that should, in principle, have identical distance from the observer?

One idea is to have the humanoid walk around a target object while maintaining the same distance from the object by keeping the perceived size the same. The main problem with this idea is that walking around an object while maintaining the perceived size of an object in the left and right image is a complicated task that would likely require high level cognitive abilities. We are interested in improving perception with simple actions. This leads us to another potential approach.

Notice that when a humanoid robot rotates its neck, the axis of rotation remains invariant, with respect to distance, to points that were not rotated. So if the robot rotates its neck, the distance from the neck to the target point remains invariant (figure 8), even though the distance from the cameras to the target may have changed. Now, if the agent estimates distance from the axis of neck rotation to the target, it can easily acquire samples that should, in principle, have identical distance to the origin.

1) Error Model: We assume the following error model

$$\begin{aligned} \tilde{\theta}_L &= \theta_L + \psi_L + \xi_L \\ \tilde{\theta}_R &= \theta_R + \psi_R + \xi_R \end{aligned} \tag{5}$$

In this error model, the angles $(\tilde{\theta}_L, \tilde{\theta}_R)$ to the target extracted from the left and right camera images are biased by ψ_L and ψ_R and corrupted by noise from the zero mean random variables ξ_L and ξ_R.

Figure 9 shows that very small biases to θ_L and θ_R can cause large errors in distance estimates as the distance of the target from the observer grows. Even a bias as small as 0.05 can cause large inaccuracies when judging the distance of targets less than two meters from the agent.

2) Estimating Distance: We assume that the learning agent has an innate algorithm for depth estimation, but the agent needs to tune several unknown parameters by learning from observations. The basic depth estimation equations (figure 10) are

$$y = \frac{\Delta}{\tan(\theta_L) + \tan(\theta_R)} \tag{6}$$

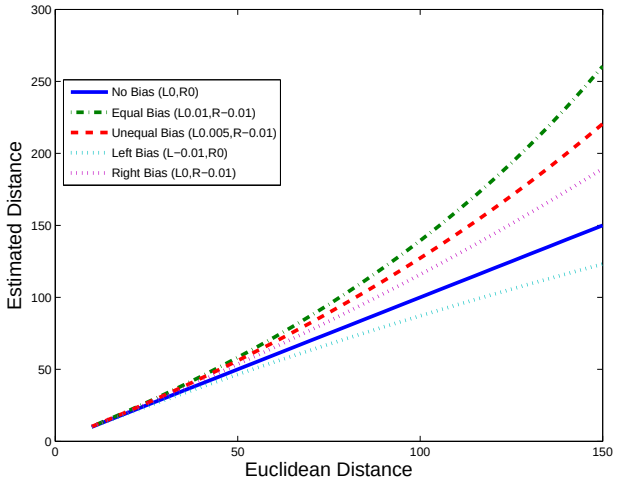


Fig. 9. Small perturbations of θ_L and θ_R cause large errors in distance estimation for far distances.

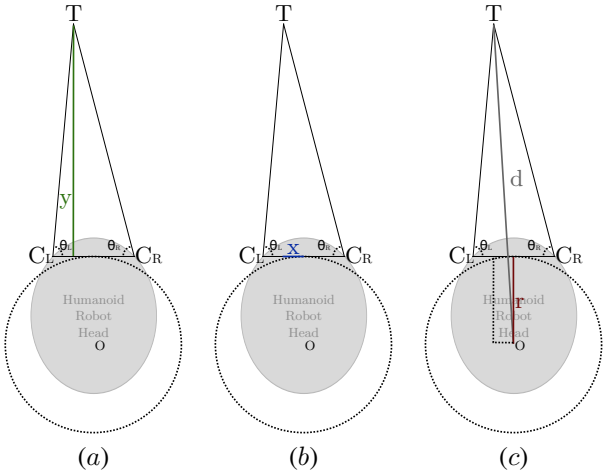


Fig. 10. Aerial view of binocular distance estimation. (a) The y quantity is the vertical distance from the agent to the target. (b) The x quantity is the horizontal distance from the point between the left and right camera and the target point. (c) The d quantity is the distance from the center of rotation to the target T, with radius r.

where y represents the axis aligned distance from the cameras to the target. Once y is computed, we can calculate

$$x = \frac{y}{\tan(\theta_L)} - \frac{\Delta}{2} \tag{7}$$

where x represents the base of the distance triangle. Finally, if the radius r is known, the distance to the axis of rotation can be computed with the following equation:

$$d = \sqrt{x^2 + (y + r)^2} \tag{8}$$

where Δ is the disparity between the left and right camera, θ_L and θ_R are the angles to the target object from the left and right cameras, respectively. Once the distances along the X-axis and Y-axis are computed (x and y, respectively), the distance d can be computed using the basic Euclidean norm. These equations are designed for the case where θ_L and θ_R are both less than $\frac{\pi}{2}$ radians. Similar equations can be derived for the other cases and are omitted for brevity.

An autonomous agent also does not know the disparity Δ between its eyes/cameras or the radius of the cameras rotation. However, examining these equations reveals that Δ effectively scales distance estimates, and so it can be arbitrarily assigned. We assume that reasonable values for Δ and r can be established by the agent using body ratio information. However, in our experiments, we provide provide appropriate values. Providing these values primarily aids in comparison with other distance estimators.

Since the agent only has access to $\tilde{\theta}_L$ and $\tilde{\theta}_R$ (and not θ_L and θ_R), the equations above will not derive the correct distance. To solve this problem, the agent must learn estimates of ψ_L and ψ_R and subtract these from $\tilde{\theta}_L$ and $\tilde{\theta}_R$ before plugging them into the distance estimation functions. Notice, however, that learning the bias terms does not mitigate the effect of noise. The best we can do is eliminate the bias.

3) Empirical Results: Learning to eliminate bias is done using a heuristic search. We generated 28 training samples Θ by rotating the robot's neck while fixating on the robot's hand. Remember that, in principle, all of the target images should have identical distance to the axis of rotation. A red circle was placed on the robot's hand to facilitate identification. The learning algorithm 1 generated $m = 1000$ random hypotheses $\left\{(\hat{\psi}_L^{(i)}, \hat{\psi}_R^{(i)})\right\}_{i=1}^{m}$. Using the training data, we evaluated each hypothesis using ~~a~~ by identifying the sum squared error of all training examples, where error was determined by how far the distance estimate fell from an arbitrary constant $c > 0$. In practice, an autonomous agent can choose any positive value for c, which determines the unit of measure that the distance estimator will use. In our case, because the target of the training samples focuses on the hand, taking $c = 1$ would result in a unit of measure that is naturally related to the robots body size. However, to facilitate comparison, we chose $c = 496$, which is the number of centimeters from the axis of neck rotation to the robot's hand.

Algorithm 1 LearnPsiBias(H, Θ, Δ, c)

```
for all h ∈ H do
    F(h) ← 0
    for all (θ̃_L, θ̃_R) ∈ Θ do
        (ψ̂_L, ψ̂_R) ← h {Extract bias hypothesis}
        d̂ ← EstimateDepth(θ̃_L − ψ̂_L, θ̃_R − ψ̂_R, Δ)
        F(h) ← F(h) + (d̂ − c)^2
    end for
end for
return arg min_{h∈H} F(h)
```

Figure 11 shows that our learning algorithm is successful at reducing the bias and improving distance estimation. The blue circles corresponding to the distance estimates of the learning agent are closer to the true distance than the estimates of the distance estimates that (incorrectly) assume zero bias.

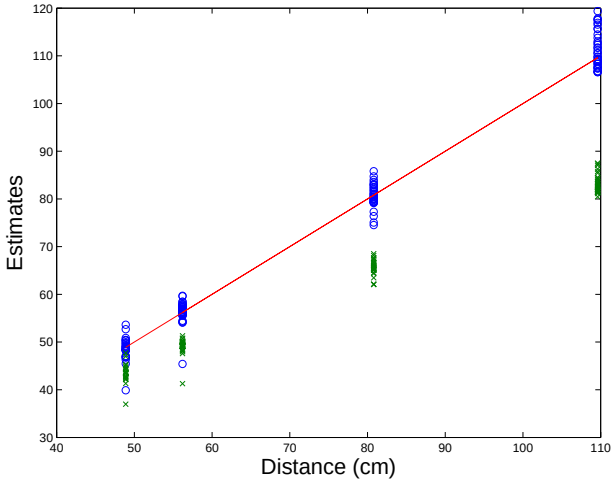


Fig. 11. Algorithm 1 corrected error introduced by bias in the estimated $\tilde{\theta}_L$ and $\tilde{\theta}_R$. Red line represents the true distance in centimeters. Blue o's denote distance predicted by the learned distance estimator. Green x's denote the distance predicted assuming $\psi_L = \psi_R = 0$.

C. Size Invariance

The perceived size of an object has an interesting relationship with distance. As an object moves closer or further from an observer, the perceived size changes, while the physical size does not. This change in perceived size is related to its change in distance from the observer. Given that a physical object typically maintains the same size (i.e. it is size invariant), we may be able to use this physical invariance to learn about distance. However, it is important to keep in mind that size itself is not, in general, a reliable clue of distance unless we know the size of the object at a reference distance. When an agent is holding an object it can easily learn the perceived size at some known reference distance and then move the object back and forth to view the object at different distances. This information can be used as a training set for our binocular vision system.

Unfortunately the relationship between perceived size and distance is not linear (see figure 12a). But there is a straightforward relationship between size and distance. We use a relationship between distance and perceived size modified from [19]

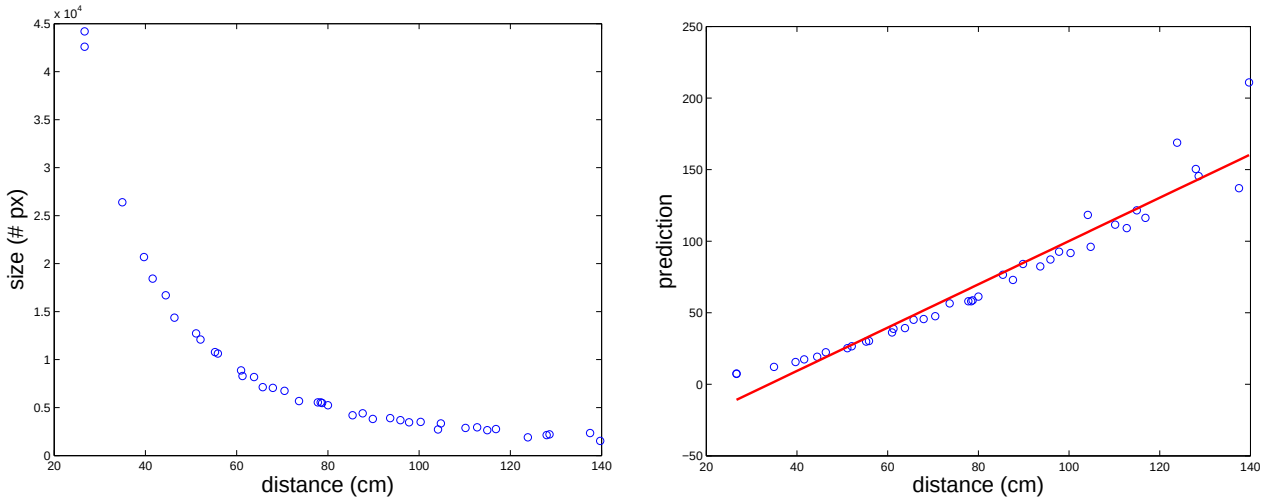
$$D = \frac{(D_0 \times s_0)}{s} + \alpha \tag{9}$$

where D is the distance given the current observation, s is perceived size of the object given the current observation, D_0 is the reference distance of the object for which the perceived size is known, s_0 is the perceived size of the object at reference distance D_0, α is a constant. Using Eq. 9, we can establish a linear relationship for training our binocular vision system (figure 12b).

The main problem with the use of size information is that the accuracy of distance estimates from size quickly degrade as distance grows. This can be seen in figure 12b. Nevertheless, size could also be useful for deriving training data for a distance estimator.

V. Human Interaction

An alternative to improving distance estimation through laborious self-experimentation is to obtain useful information



(a) (b)

Fig. 12. (a) The relationship between perceived size and distance of an object is a nonlinear inverted relationship. (b) However, the perceived size/distance relationship can be reliably converted to a linearly correlated relationship.

from other agents. However, relying completely on other agents to acquire information places too much of the burden on other agents. Instead, we are interested in improving distance estimation with very limited but useful information given from humans. In this case, the information being given to the robot is its distance from a target. We consider the case where the distance to just one or two targets are given to the robot.

Another interesting possibility is that the robot can receive information that it might not otherwise be capable of obtaining on its own. For example, the robotic system described in the previous section only learned about distance from a single target on its own body.

In this experiment, we compared distance estimates of two types of agents. The first learning system learned depth autonomously (as in the previous section) by choosing a single target (at 40cm) and rotating its neck. In this way, the agent could generate many samples to train on, but they were all for the same target object - the robot's hand. Thus all training samples were 40cm from the robot. The second learning system used the a similar strategy, but in addition to its autonomously selected target point (at 40cm), the robot was told the distance to one additional target (at 1600cm) by a human. With the second target's distance the robot was able to generate even more samples by rotating its neck.

In both cases, bias parameters $\hat{\psi}_L$ and $\hat{\psi}_R$ were learned using algorithm 1 with 1,000 randomly generated hypotheses. Figure 13 shows that for distant objects, the agents that were given information from a human generated superior hypotheses. This experiment was repeated nine times for the autonomous learning case and three times for autonomous learning with human input to generate estimates of the standard deviations.

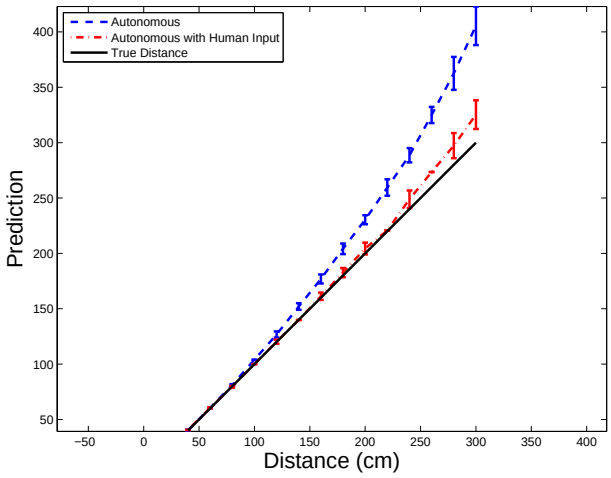


Fig. 13. Comparison of distance estimates from an autonomous distance estimator, autonomous distance estimator given one additional sample from a human, and the true distance. For long distances, the autonomous distance estimator with human input is more accurate. Error bars depict standard deviations.

VI. Discussion

The main contributions of this work is the demonstration of how action and interaction can be applied to improving perception. We have specifically demonstrated these abilities on the egocentric distance estimation task. By exploiting invariant properties, an autonomous system can expose inconsistencies in its perceptual processing. In the case of neck rotation, the action leaves the true physical distance unaltered and any disagreement between estimates provides an error signal for training. In the case of perceived size, the physical size of the object remains invariant, but the perceived size changes. Since we can establish a linear relationship between a function of perceived size and distance, we can use perceived size of a training target object, such as the robot's hand, to learn appropriate parameters for the general, egocentric distance estimation using the binocular vision system.

Interestingly, we see in figure 13 that a single input from a human can further improve distance estimation at distant targets. This small level of effort by the human is made

possible by the robot taking full advantage of the single additional piece of information, again through action.

One limitation of this work is that the error model (Eq. 5) is quite simple. Although our algorithms were able to reduce distance estimation error, better results may be possible by assuming a more complex error model. For example, the bias to θ_L and θ_R may increase as the the target object moves into the peripheries of the image. More complex error models may offer an interesting line of future investigation. These error models may also incorporate lower level parameters used to derive $\tilde{\theta}_L$ and $\tilde{\theta}_R$ from image data, such as the focal length.

However, these more complex error models may not be learnable with our simple neck rotation strategy. This leads to another important problem: integration of multiple cues for depth. By incorporating size information and neck rotation, as well as other strategies, it is possible to constrain more complex error models and as a result produce better distance estimates. Investigating integration of distance estimation cues is an interesting area of further investigation.

Another interesting direction of future work would be to investigate the role of action for estimating distance using a motion parallax strategy. The method is theoretically quite similar to that used for binocular distance estimation, but motion parallax is more strongly coupled with action and can be implemented on a system with only a single camera or used in combination with binocular distance estimation to obtain more accurate distance estimates.

VII. Conclusion

We have demonstrated that both action and interaction with social agents are valuable methods for an embodied autonomous agent to improve its perception. Actions that maintain perceptual invariance can be used to measure the error of a perceptual process, such as distance estimation. Interaction with social agents can be used to acquire information that is difficult to obtain by oneself. Together these techniques form a valuable toolkit for improving perception.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. 1107652.

References

[1] E. G. O. de Haart, D. P. Carey, and A. B. Milne, "More thoughts on perceiving and grasping the müller-lyer illusion," *Neuropsychologia*, vol. 37, pp. 1437–1444, 1999.

[2] S. Aglioti, J. F. DeSouza, and M. A. Goodale, "Size-contrast illusions deceive the eye but not the hand," *Current Biology*, vol. 5, pp. 679–685, 1995.

[3] R. Held and A. Hein, "Movement-produced stimulation in the development of visually guided behavior," *Journal of Comparative and Physiological Psychology*, vol. 56, pp. 872–876, 1963.

[4] M. Wexler and J. J. van Broxtel, "Depth perception by the active observer," *Trends in Cognitive Sciences*, vol. 9, no. 9, pp. 431–438, 2005.

[5] B. Calvo-Merino, D. Glaser, J. Grèzes, R. Passingham, and P. Haggard, "Action observation and acquired motor skills: An fMRI study of expert dancers," *Cerebral Cortex*, vol. 15, pp. 1243–1249, 2005.

[6] T. T.-J. Chong, R. Cunnington, M. A. Williams, N. Kanwisher, and J. B. Mattingley, "fmri adaptation reveals mirror neurons in human inferior parietal cortex," *Current Biology*, vol. 18, no. 20, pp. 1576 – 1580, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960982208012426

[7] V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti, "Action recognition in the permotor cortex," *Brain*, vol. 119, pp. 593–609, 1996.

[8] B. Julesz, "Binocular depth perception without familiarity cues," *Science*, vol. 145, no. 3630, pp. 356–362, 1964. [Online]. Available: http://www.sciencemag.org/content/145/3630/356.abstract

[9] G. F. Poggio and T. Poggio, "The analysis of stereopsis," *Annual Review of Neuroscience*, vol. 7, pp. 379–412, 1984.

[10] B. Rogers and M. Graham, "Similarities between motion parallax and stereopsis in human depth perception," *Vision Research*, vol. 22, pp. 261–270, 1982.

[11] S.-B. Choi, B.-S. Jung, S.-W. Ban, H. Niitsuma, and M. Lee, "Biologically motivated vergence control system using human-like selective attention model," *Neurocomputing*, vol. 69, pp. 537–558, 2006.

[12] S. Jeong, S.-W. Ban, and M. Lee, "Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment," *Neural Networks*, vol. 21, no. 10, pp. 1420–1430, 2008.

[13] F. Thorn, J. Gwiazda, A. A. V. Cruz, J. A. Bauer, and R. Held, "The development of eye alignment, convergence, and sensory binocularity in young infants," *Investigative Ophthalmology and Visual Science*, vol. 35, pp. 544–553, 1994.

[14] A. M. Treisman and G. Gelde, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[15] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[16] S.-J. Park, K.-H. An, and M. Lee, "Saliency map model with adaptive masking based on independent component analysis," *Neurocomputing*, vol. 49, no. 1, pp. 417–422, 2002.

[17] T. Kadir and M. Brady, "Scale, saliency and image description," *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.

[18] Y. Choe, H.-F. Yang, and D. C.-Y. Eng, "Autonomous learning of the semantics of internal sensory states based on motor exploration," *International Journal of Humanoid Robotics*, vol. 4, pp. 211–243, 2007.

[19] A. S. Gilinsky, "Perceived size and distance in visual space," *Psychological Review*, vol. 58, pp. 460–482, 1951.