[P2-15]

Autonomously improving binocular depth estimation

Timothy A. Mann[1]*, Yunjung Park[2], Sungmoon Jeong[2], Minho Lee (P)[2], and Yoonsuck Choe[1]

[1] Department of Computer Science & Engineering, Texas A&M University
[2] School of Electrical Engineering and Computer Science, Kyungpook National University
E-mail: mann23@tamu.edu,yj-park@ee.knu.ac.kr,jeongsm@ee.knu.ac.kr,mholee@knu.ac.kr,choe@tamu.edu

Abstract— We investigate how an autonomous humanoid robot with an initially inaccurate binocular vision system can learn to correct inconsistencies in its understanding of distance using information and resources that might be available to a human infant. We defined a consistent depth estimator as a Euclidean distance metric where the unit of measurement is determined autonomously. We found that an error signal that exploits actions that maintain invariant distance is a powerful tool for correcting inconsistency. Our results show that a heuristic search algorithm, run incrementally as new data become available, can efficiently (i.e. with few samples) correct inconsistencies and improve depth estimates.

Keywords— Binocular Vision, Depth Estimation, Active Learning, Autonomous Robot

1 Problem

Because humans and most animals are born with two eyes, they have the necessary "hardware" to estimate the egocentric distance to physical objects. Species with two eyes are also likely to have evolved innate or developmental mechanisms for depth estimation because the algorithms for estimating depth are constant except for the values of a few parameters. The main problem faced by these embodied agents is selecting appropriate parameters and adjusting sensory input to match the assumptions made by its innate depth estimation mechanism and distance in the environment. The main question addressed by this study is: How can an autonomous embodied agent with an initially inaccurate binocular vision system learn to correct inconsistencies in its understanding of distance using information and resources that might be available to a human infant?

2 Method

In our experiments, a distance estimator was considered to be consistent if the expected value of its estimates are proportional to Euclidean distance in our preferred unit of measure (e.g. centimeters). Formally, a consistent distance estimator satisfies

$$E\left[\hat{d}\right] = \alpha D \tag{1}$$

where $\hat{d}$ is the distance estimate, $E[\cdot]$ is the expectation operator, $\alpha \in \mathbb{R}^+$ is a positive scalar, and D is the distance in our preferred unit of measure.

*This research has been supported in part by the NSF East Asia Pacific Summer Institutes (EAPSI) program.

Figure 1: A depth estimation problem where C_L and C_R represent the left and right cameras, respectively, and T is the target. Distance d can be derived from the angles θ_L and θ_R and the disparity Δ between the cameras.

The basic depth estimation equations (figure 1) are

$$\begin{aligned} y &= \frac{\Delta}{\tan(\theta_L)+\tan(\theta_R)} \\ x &= \frac{y}{\tan(\theta_L)} - \frac{\Delta}{2} \\ d &= \sqrt{x^2+(y+r)^2} \end{aligned} \tag{2}$$

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.

Equation (2) assumes that the given θ_L, θ_R, Δ, and r are accurate. However, an autonomous agent may estimate θ_L or θ_R inaccurately because it does not know the correct properties of its eyes/cameras. An autonomous agent also does not know the disparity Δ between its eyes/cameras or the radius of the cameras rotation. However, examining Eq. (2) 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. θ_L and θ_R on the other hand must be accurate, or else the agent's depth estimates will be inconsistent. We consider a simple bias model for the angles observed by the embodied agent:

$$\begin{aligned} \tilde{\theta}_L &= \theta_L + \beta_L + \epsilon_L \\ \tilde{\theta}_R &= \theta_R + \beta_R + \epsilon_R \end{aligned} \tag{3}$$

where $\tilde{\theta}_L$ and $\tilde{\theta}_R$ are angles observed by the agent, θ_L and θ_R are the true angles, β_L and β_R are bias terms, and ϵ_L and ϵ_R are Gaussian noise. Under this model, it is sufficient for the agent to learn β_L and β_R so that it can correct for this bias.

We designed an error signal based on the idea that distance is invariant under some actions. Our idea here is inspired by [1] where the objective is to maintain sensory invariance to solve a similar problem where there is only an internal observer and no external teacher. Here we specifically considered rotation of our humanoid robot's neck. If the origin is defined at the

Figure 2: Nao robot with mounted cameras.

axis of rotation (see figure 1), then the target objects distance to the origin will remain invariant during rotation. Our error signal exploits the embodiment of our robot. We assume that the robot can move to a reference pose where its hand is visible. In this reference pose, the agent can arbitrarily assign the egocentric distance to its hand. The assigned distance then determines the agent's unit of measurement. The error signal used in the experiments was

$$\xi = \tanh\left((\hat{d} - c)^2\right) \tag{4}$$

where ξ is the error, $\hat{d}$ is the estimated distance, and c is the distance assigned to the reference pose. The hyperbolic tangent function $h(\cdot)$ was only used to prevent numerical instability. This signal combined with neck rotation reveals depth estimation inconsistencies.

Algorithm 1 EvaluatePopulation(H, Θ, Δ, c)

```
for all h ∈ H do
  F(h) ← 0
  for all (θ̃_L, θ̃_R) ∈ Θ do
    (β̂_L, β̂_R) ← h {Extract bias from hypothesis}
    d̂ ← EstimateDepth(θ̃_L − β̂_L, θ̃_R − β̂_R, Δ)
    F(h) ← F(h) + tanh((d̂ − c)^2)
  end for
end for
return F
```

We used a heuristic search algorithm to identify the best hypothesis. Algorithm 1 describes how a set of hypotheses can be evaluated using Eq. (4). For each hypothesis $h = \left(\hat{\beta}_L, \hat{\beta}_R\right) \in H$ and each data sample in Θ, the algorithm sums up the error in $F(h)$. In the fitness array $F(\cdot)$, the best hypotheses have low values.

3 Experiment & Results

We experimented with an Aldebaran Nao humanoid robot with two cameras mounted on its head (figure 2). The correspondence problem between the left and right camera was solved using the visual attention system described in [2]. A green circle was placed on the end of the robot's arm to facilitate tracking.

Images from the left and right cameras were sampled as the robot rotated its neck and converted to a saliency map, used to extract the center of the hand or target. We systematically added bias to samples of θ_L and θ_R to simulate the effect of an inconsistent distance estimator. The computational complexity of our algorithm is linear in the number of hypotheses and the number of image samples. Only 1,000 total hypotheses were generated and evaluated using algorithm 1, because the hypothesis space is low dimensional and reasonable solutions were identified without further search.

Figure 3 compares an initially inconsistent distance estimator with the learned distance estimator plotted against distance in centimeters. These results are generated by the simulator so that the effect of noise can be removed and many target distances can be evaluated but the hypotheses were evaluated against actual data from the robot. Notice that the learned distance estimator is almost directly proportional to the distance in centimeters achieving our objective of learning a Euclidean distance estimator (Eq. (1)).

Figure 3: Comparison between original inconsistent distance estimator and learned distance estimator plotted against distance in centimeters.

4 Discussion & Conclusion

Learning to maintain perceptual invariance of the distance while rotating the neck corrects inconsistent distance estimation. First, what is striking about our method is that the error signal can be generated using only simple actions, such as rotating the neck, and low level visual features. The embodied agent does not need highly coordinated actions such as walking or manipulating objects. Second, by evaluating multiple hypotheses with each data sample, the learning mechanism needs little data to select the best hypotheses compared to gradient based methods, which need many samples per iteration.

References

[1] Yoonsuck Choe, Huei-Fang Yang, and Daniel Chern-Yeow Eng. Autonomous learning of the semantics of internal sensory states based on motor exploration. *International Journal of Humanoid Robotics*, 4:211–243, 2007.

[2] Sungmoon Jeong, Sand-Woo Ban, and Minho Lee. Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment. *Neural Networks*, 21:1420–1430, 2008.