



Peripheral-Foveal Vision for Real-time Object Recognition



Stephen Gould, Benjamin Sapp, Morgan Quigley, Andrew Y. Ng

1 Overview

- Human object recognition in a 3d environment is far superior to that of any robotic vision system.
- One reason (out of many) for this is that humans use a **fovea** to fixate on, or near an object, thus obtaining a very high resolution image of the object and rendering it easy to recognize.
- We present a novel method for identifying and tracking objects using a two camera system.
- Our method** is motivated by biological vision systems:
 - uses a learned “**attentive**” **interest map** on a low resolution view to direct a high resolution “**fovea**.”
 - objects that are recognized in the **fovea** can then be tracked using **peripheral vision**.
- Object recognition is run only on a small foveal image improving real-time performance.
- Our work is related to [Orabona *et al.*, 2005], [Tagare *et al.*, 2001], and [Ude *et al.*, 2003].

2 STAIR robot



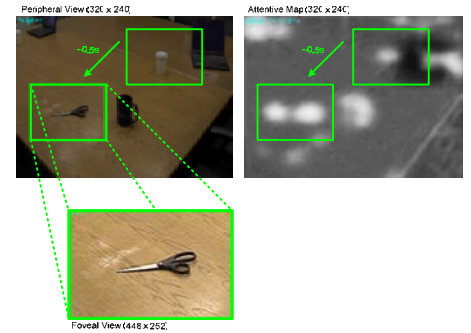
- The **STAIR** (STanford Artificial Intelligence Robot) project has long-term goal of integrating techniques from all areas of AI to build a useful home/office assistant robot.
- Goal of this work: **object recognition system for STAIR.**

Acknowledgements

We would like to thank Adrian Kaehler and Gary Bradski for their contributions to this work.

3 Visual attention model

- Our system uses two separate cameras:
 - fixed **wide-angle camera** for **peripheral vision**
 - controllable **pan-tilt-zoom (PTZ)** camera for **foveal vision**.
- The PTZ camera can focus on any region of the scene to obtain a high-resolution image for object recognition.
- Previously identified objects are tracked using peripheral vision.
- The attention system periodically decides between the following actions:
 - Confirmation** of a tracked object by fixating the fovea over the predicted location of the object;
 - Search** for unidentified objects by moving the fovea to some new part of the scene.
- Estimating the reduction in entropy, H , by taking each action (**Confirmation** or **Search**), we take the action which maximizes the expected reduction in entropy.



- Let $\xi_k(t)$ denote the state of the k -th object, α_k , at time t . Then (assuming independence of objects) we can compute our uncertainty over objects in the scene as,

$$H = \sum_{\alpha_k \in \mathcal{I}} H_{\text{tracked}}(\xi_k(t)) + \sum_{\alpha_k \notin \mathcal{I}} H_{\text{unidentified}}(\xi_k(t))$$

- $H_{\text{untracked}}(\xi_k(t))$ cannot be computed directly since we do not know the number of objects in the scene so we use an **interest model** to estimate the probability of finding an unknown object in a given foveal region.

4 Interest modeling

- Our **interest model** allows us to choose which foveal region to examine next by rapidly identifying pixels which have a high probability of containing objects that we can classify.
- We define a pixel to be interesting if it is **part of an unknown, yet classifiable object**.
 - a consequence of this definition is that our model automatically encodes the biological phenomena of **saliency** and **inhibition of return** [Itti and Koch, 2001].
- Interestingness** of every pixel in the peripheral view is modeled using a dynamic Bayesian network (DBN) whose parameters are learned from training videos.

6 Experimental results

- Our method compared to three naive approaches:
 - fixing the foveal gaze to the center of view,
 - linearly scanning over the scene from top-left to bottom-right, and,
 - randomly moving the fovea around the scene.

| Fovea control | Recall | Precision | F ₁ -Score |
|-------------------|--------------|--------------|-----------------------|
| Fixed at center | 9.49% | 97.4% | 17.3% |
| Linear scanning | 13.6% | 100.0% | 24.0% |
| Random scanning | 27.7% | 84.1% | 41.6% |
| Our method | 62.2% | 83.9% | 71.5% |

- Videos demonstrating our results are available at <http://ai.stanford.edu/~sgould/vision/>

5 Object recognition and tracking

- A **Kalman filter** tracks the location and velocity of identified objects in the 2d image plane.
- Use subset of biologically inspired **C1 features** [Serre *et al.*, 2004] and learn a boosted decision tree classifier for each object.



References

- Peripheral-foveal vision for real-time object recognition and tracking in video**, Stephen Gould, Joakim Arfvidsson, Adrian Kaehler, Benjamin Sapp, Marius Meissner, Gary Bradski, Paul Baumstarck, Sukwon Chung and Andrew Y. Ng. To appear in *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)*, 2007.
- Computational modeling of visual attention**, L. Itti and C. Koch. *Nature Reviews Neuroscience*, 2(3):194–203, Mar 2001.
- A new biologically motivated framework for robust object recognition**, Thomas Serre, Lior Wolf, and Tomaso Poggio. In *AI Memo 2004-026*, November 2004.