Peripheral-Foveal Vision for Real-time Object Recognition and Tracking in Video

Stephen Gould, Joakim Arfvidsson, Adrian Kaehler, Benjamin Sapp, Marius Messner, Gary Bradski, Paul Baumstarck, Sukwon Chung and Andrew Y. Ng
Stanford University
Stanford, CA 94305 USA

Abstract

Human object recognition in a physical 3-d environment is still far superior to that of any robotic vision system. We believe that one reason (out of many) for this—one that has not heretofore been significantly exploited in the artificial vision literature—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. In this paper, we present a novel method for identifying and tracking objects in multi-resolution digital video of partially cluttered environments. Our method is motivated by biological vision systems and uses a learned "attentive" interest map on a low resolution data stream to direct a high resolution "fovea." Objects that are recognized in the fovea can then be tracked using peripheral vision. Because object recognition is run only on a small foveal image, our system achieves performance in real-time object recognition and tracking that is well beyond simpler systems.

1 Introduction

The human visual system far outperforms any robotic system in terms of object recognition. There are many reasons for this, and in this paper, we focus only on one hypothesis—which has heretofore been little exploited in the computer vision literature—that the division of the retina into the foveal and peripheral regions is of key importance to improving performance of computer vision systems on continuous video sequences. Briefly, the density of photoreceptor rod and cone cells varies across the retina. The small central fovea, with a high density of color sensitive cone cells, is responsible for detailed vision, while the surrounding periphery, containing a significantly lower density of cones and a large number of monochromatic rod cells, is responsible for, among other things, detecting motion. Estimates of the equivalent number of "pixels" in the human eye vary, but based on spatial acuity of $1/120$ degrees [Edelman and Weiss, 1995; Fahle and Poggio, 1981], appears to be on the order of 5×10^8 pixels. For a more in-depth treatment of human visual perception, we refer the reader to one of the many excellent textbooks in the literature, for example [Wandell, 1995].

Because very little changes in a visual stream from one frame to the next, one expects that it is possible to identify objects or portions of a scene one at a time using the high resolution fovea, while tracking previously identified objects in the peripheral region. The result is that it is possible to use computationally expensive classification algorithms on the relatively limited portion of the scene on which the fovea is fixated, while accumulating successful classifications over a series of frames in real-time.

A great deal has happened in recent years in the area of location and identification of specific objects or object classes in images. Much of this work, however, has concentrated on single frames with the object of interest taking up a significant proportion of the field-of-view. This allows for accurate and robust object recognition, but implies that if we wish to be able to find very small objects, we must go to much higher image resolutions.[1] In the naive approach, there is significant computational cost of going to this higher resolution.

For continuous video sequences, the standard technique is simply to treat each frame as a still image, run a classification algorithm over the frame, and then move onto the next frame, typically using overlap to indicate that an object found in frame $n + 1$ is the same object found in frame n, and using a Kalman filter to stabilize these measurements over time. The primary difficulty that this simple method presents is that a tremendous amount of effort is misdirected at the vast majority of the scene which does not contain the objects we are attempting to locate. Even if the Kalman filter is used to predict the new location of an object of interest, there is no way to detect the appearance of new objects which either enter the scene, or which are approached by a camera in motion.

The solution we propose is to introduce a peripheral-foveal model in which attention is directed to a small portion of the visual field. We propose using a low-resolution, wide-angle video camera for peripheral vision and a pan-tilt-zoom (PTZ) camera for foveal vision. The PTZ allows very high resolution on the small portion of the scene in which we are interested at any given time. We refer to the image of the scene supplied by the low-resolution, wide-angle camera as the *peripheral view* and the image supplied by the pan-tilt-zoom

[1] Consider, for example, a typical coffee mug at a distance of five meters appearing in a standard 320×240 resolution, 42° field-of-view camera. The 95mm tall coffee mug is reduced to a mere 8 pixels high, rendering it extremely difficult to recognize.

camera as the *foveal view* or simply *fovea*.

We use an attentive model to determine regions from the peripheral view on which we wish to perform object classification. The attentive model is learned from labeled data, and can be interpreted as a map indicating the probability that any particular pixel is part of an unidentified object. The fovea is then repeatedly directed by choosing a region to examine based on the expected reduction in our uncertainty about the location of objects in the scene. Identification of objects in the foveal view can be performed using any of the state-of-the-art object classification technologies (for example see [Viola and Jones, 2004; Serre *et al.*, 2004; Brubaker *et al.*, 2005]).

The PTZ camera used in real-world robotic applications can be modeled for study (with less peak magnification) using a high-definition video (HDV) camera, with the "foveal" region selected from the video stream synthetically by extracting a small region of the HDV image. The advantage of this second method is that the input data is exactly reproducible. Thus we are able to evaluate different foveal camera motions on the same recorded video stream. Figure 1 shows our robot mounted with our two different camera setups.

Figure 1: STAIR platform (left) includes a low-resolution peripheral camera and high-resolution PTZ camera (top-right), and alternative setup with HDV camera replacing the PTZ (bottom-right). In our experiments we mount the cameras on a standard tripod instead of using the robotic platform.

The robot is part of the STAIR (STanford Artificial Intelligence Robot) project, which has the long-term goal of building an intelligent home/office assistant that can carry out tasks such as cleaning a room after a dinner party, and finding and fetching items from around the home or office. The ability to visually scan its environment and quickly identify objects is of one of the key elements needed for a robot to accomplish such tasks.

The rest of this paper is organized as follows. Section 2 outlines related work. In Section 3 we present the probabilistic framework for our attention model which directs the foveal gaze. Experimental results from a sample HDV video stream as well as real dual-camera hardware are presented in Section 4. The video streams are of a typical office environment and contain distractors, as well as objects of various types for which we had previously trained classifiers. We show that the performance of our attention driven system is significantly better than naive scanning approaches.

2 Related Work

An early computational architecture for explaining the visual attention system was proposed by Koch and Ullman (1985) . In their model, a scene is analyzed to produce multiple feature maps which are combined to form a saliency map. The single saliency map is then used to bias attention to regions of highest activity. Many other researchers, for example [Hu *et al.*, 2004; Privitera and Stark, 2000; Itti *et al.*, 1998], have suggested adjustments and improvements to Koch and Ullman's model. A review of the various techniques is presented in [Itti and Koch, 2001].

A number of systems, inspired by both physiological models and available technologies, have been proposed for finding objects in a scene based on the idea of visual attention, for example [Tagare *et al.*, 2001] propose a maximum-likelihood decision strategy for finding a known object in an image.

More recently, active vision systems have been proposed for robotic platforms. Instead of viewing a static image of the world, these systems actively change the field-of-view to obtain more accurate results. A humanoid system that detects and then follows a single known object using peripheral and foveal vision is presented in [Ude *et al.*, 2003]. A similar hardware system to theirs is proposed in [Björkman and Kragic, 2004] for the task of object recognition and pose estimation. And in [Orabona *et al.*, 2005] an attention driven system is described that directs visual exploration for the most salient object in a static scene.

An analysis of the relationship between corresponding points in the peripheral and foveal views is presented in [Ude *et al.*, 2006], which also describes how to physically control foveal gaze for rigidly connected binocular cameras.

3 Visual Attention Model

Our visual system comprises two separate video cameras. A fixed wide-angle camera provides a continuous low-resolution video stream that constitutes the robot's peripheral view of a scene, and a controllable pan-tilt-zoom (PTZ) camera provides the robot's foveal view. The PTZ camera can be commanded to focus on any region within the peripheral view to obtain a detailed high-resolution image of that area of the scene. As outlined above, we conduct some of our experiments on recorded high-resolution video streams to allow for repeatability in comparing different algorithms.

Figure 2 shows an example of a peripheral and foveal view from a typical office scene. The attention system selects a foveal window from the peripheral view. The corresponding region from the high-resolution video stream is then used for

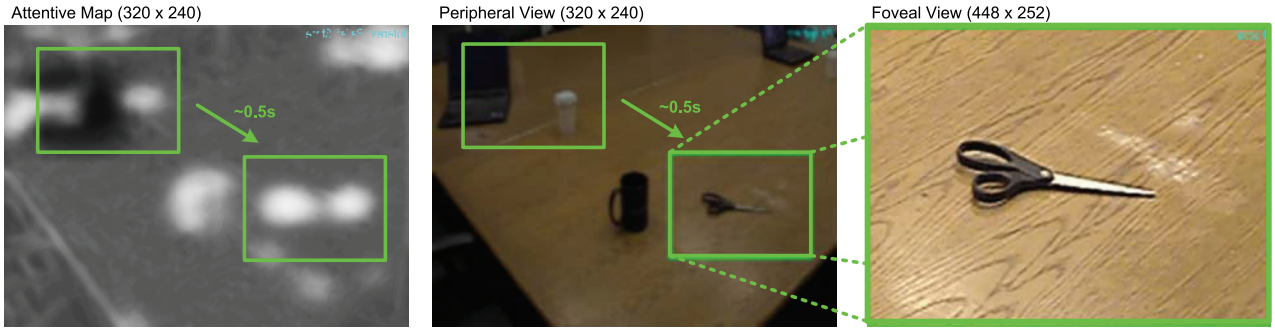

Figure 2: Illustration of the peripheral (middle) and foveal (right) views of a scene in our system with attentive map showing regions of high interest (left). In our system it takes approximately 0.5 seconds for the PTZ camera to move to a new location and acquire an image.

classification. The attentive interest map, generated from features extracted from the peripheral view, is used to determine where to direct the foveal gaze, as we will discuss below.

The primary goal of our system is to identify and track objects over time. In particular, we would like to minimize our uncertainty in the location of all identifiable objects in the scene in a computationally efficient way. Therefore, the attention system should select regions of the scene which are most informative for understanding the robot's visual environment.

Our system is able to track previously identified objects over consecutive frames using peripheral vision, but can only classify new objects when they appear in the (high-resolution) fovea, $\mathcal{F}$. Our uncertainty in a tracked object's position grows with time. Directing the fovea over the expected position of the object and re-classifying allows us to update our estimate of its location. Alternatively, directing the fovea to a different part of the scene allows us to find new objects. Thus, since the fovea cannot move instantaneously, the attention system needs to periodically decide between the following actions:

A1. Confirmation of a tracked object by fixating the fovea over the predicted location of the object to confirm its presence and update our estimate of its position;

A2. Search for unidentified objects by moving the fovea to some new part of the scene and running the object classifiers over that region.

Once the decision is made, it takes approximately 0.5 seconds—limited by the speed of the PTZ camera—for the fovea to move to and acquire an image of the new region.[2] During this time we track already identified objects using peripheral vision. After the fovea is in position we search for new and tracked objects by scanning over all scales and shifts within the foveal view as is standard practice for many state-of-the-art object classifiers.

More formally, let $\xi_k(t)$ denote the state of the k-th object, o_k, at time t. We assume independence of all objects in the scene, so our uncertainty is simply the sum of entropy terms over all objects,

$$H = \sum_{o_k \in \mathcal{T}} H_{\text{tracked}}(\xi_k(t)) + \sum_{o_k \notin \mathcal{T}} H_{\text{unidentified}}(\xi_k(t)) \quad (1)$$

where the first summation is over objects being tracked, $\mathcal{T}$, and the second summation is over objects in the scene that have not yet been identified (and therefore are not being tracked).

Since our system cannot know the total number of objects in the scene, we cannot directly compute the entropy over unidentified objects, $\sum_{o_k \notin \mathcal{T}} H(\xi_k(t))$. Instead, we learn the probability $P(o_k \mid \mathcal{F})$ of finding a previously-unknown object in a given foveal region $\mathcal{F}$ of the scene based on features extracted from the peripheral view (see the description of our interest model in section 3.2 below). If we detect a new object, then one of the terms in the rightmost sum of Eq. (1) is reduced; the expected reduction in our entropy upon taking action **A2** (and fixating on a region $\mathcal{F}$) is

$$\Delta H_{A2} = P(o_k \mid \mathcal{F})\big(H_{\text{unidentified}}(\xi_k(t)) - H_{\text{tracked}}(\xi_k(t+1))\big). \quad (2)$$

Here the term $H_{\text{unidentified}}(\xi_k(t))$ is a constant that reflects our uncertainty in the state of untracked objects,[3] and $H_{\text{tracked}}(\xi_k(t+1))$ is the entropy associated with the Kalman filter that is attached to the object once it has been detected (see Section 3.1).

As objects are tracked in peripheral vision, our uncertainty in their location grows. We can reduce this uncertainty by observing the object's position through re-classification of the area around its expected location. We use a Kalman filter to track the object, so we can easily compute the reduction in entropy from taking action **A1** for any object $o_k \in \mathcal{T}$,

$$\Delta H_{A1} = \tfrac{1}{2}\left(\ln|\Sigma_k(t)| - \ln|\Sigma_k(t+1)|\right) \quad (3)$$

where $\Sigma_k(t)$ and $\Sigma_k(t+1)$ are the covariance matrices associated with the Kalman filter for the k-th object before and after re-classifying the object, respectively.

[2] In our experiments with single high-resolution video stream, we simulate the time required to move the fovea by delaying the image selected from the HDV camera by the required number of frames.

[3] Our experiments estimated this term assuming a large-variance distribution over the peripheral view of possible object locations; in practice, the algorithm's performance appeared very insensitive to this parameter.

In this formalism, we see that by taking action **A1** we reduce our uncertainty in a tracked object's position, whereas by taking action **A2** we may reduce our uncertainty over unidentified objects. We assume equal costs for each action and therefore we choose the action which maximizes the expected reduction of the entropy H defined in Eq. (1).

We now describe our tracking and interest models in detail.

3.1 Object Tracking

We track identified objects using a Kalman filter. Because we assume independence of objects, we associate a separate Kalman filter with each object. An accurate dynamic model of objects is outside the current scope of our system, and we use simplifying assumptions to track each object's coordinates in the 2-d image plane. The state of the k-th tracked object is

$$\xi_k = [x_k \quad y_k \quad \dot{x}_k \quad \dot{y}_k]^T \tag{4}$$

and represents the (x, y)-coordinates and (x, y)-velocity of the object.

On each frame we perform a Kalman filter motion update step using the current estimate of the object's state (position and velocity). The update step is

$$\xi_k(t+1) = \begin{bmatrix} 1 & 0 & \Delta T & 0 \\ 0 & 1 & 0 & \Delta T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \xi_k(t) + \eta_m \tag{5}$$

where $\eta_m \sim \mathcal{N}(0, \Sigma_m)$ is the motion model noise, and ΔT is the duration of one timestep.

We compute optical flow vectors in the peripheral view generated by the Lucas and Kanade (1981) sparse optical flow algorithm. From these flow vectors we measure the velocity of each tracked object, $z_v(t)$, by averaging the optical flow within the object's bounding box. We then perform a Kalman filter observation update, assuming a velocity observation measurement model

$$z_v(t) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \xi_k(t) + \eta_v \tag{6}$$

where $\eta_v \sim \mathcal{N}(0, \Sigma_v)$ is the velocity measurement model noise.

When the object appears in the foveal view, i.e., after taking action **A1**, the classification system returns an estimate, $z_p(t)$, of the object's position, which we use to perform a corresponding Kalman filter observation update to greatly reduce our uncertainty in its location accumulated during tracking. Our position measurement observation model is given by

$$z_p(t) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \xi_k(t) + \eta_p \tag{7}$$

where $\eta_p \sim \mathcal{N}(0, \Sigma_p)$ is the position measurement model noise.

We incorporate into the position measurement model the special case of not being able to re-classify the object even though the object is expected to appear in the foveal view. For example, this may be due to misclassification error, poor estimation of the object's state, or occlusion by another object. In this case we assume that we have lost track of the object.

Since objects are recognized in the foveal view, but are tracked in the peripheral view, we need to transform coordinates between the two views. Using the well-known stereo calibration technique of [Zhang, 2000], we compute the extrinsic parameters of the PTZ camera with respect to the wide-angle camera. Given that our objects are far away relative to the baseline between the cameras, we found that the disparity between corresponding pixels of the objects in each view is small and can be corrected by local correlation.[4] This approach provides adequate accuracy for controlling the fovea and finding the corresponding location of objects between views.

Finally, the entropy of a tracked object's state is required for deciding between actions. Since the state, $\xi(t)$, is a Gaussian random vector, it has differential entropy

$$H_{\text{tracked}}(\xi_k(t)) = \tfrac{1}{2} \ln(2\pi e)^4 |\Sigma_k(t)| \tag{8}$$

where $\Sigma_k(t) \in \mathbb{R}^{4\times4}$ is the covariance matrix associated with our estimate of the k-th object at time t.

3.2 Interest Model

To choose which foveal region to examine next (under action **A2**), we need a way to estimate the probability of detecting a previously unknown object in any region $\mathcal{F}$ in the scene. To do so, we define an "interest model" that rapidly identifies pixels which have a high probability of containing objects that we can classify. A useful consequence of this definition is that our model automatically encodes the biological phenomena of *saliency* and *inhibition of return*—the processes by which regions of high visual stimuli are selected for further investigation, and by which recently attended regions are prevented from being attended again (see [Klein, 2000] for a detailed review).

In our approach, for each pixel in our peripheral view, we estimate the probability of it belonging to an unidentified object. We then build up a map of regions in the scene where the density of interest is high. From this interest map our attention system determines where to direct the fovea to achieve maximum benefit as described above.

More formally, *we define a pixel to be interesting if it is part of an unknown, yet classifiable object*. Here *classifiable*, means that the object belongs to one of the object classes that our classification system has been trained to recognize. We model interestingness, $\mathbf{y}(t)$, of every pixel in the peripheral view, $\mathbf{x}(t)$, at time t using a dynamic Bayesian network (DBN), a fragment of which is shown in Figure 3. For the (i, j)-th pixel, $y_{i,j}(t) = 1$ if that pixel is interesting at time t, and 0 otherwise. Each pixel also has associated with it a vector of observed features $\phi_{i,j}(\mathbf{x}(t)) \in \mathbb{R}^n$.

[4] We resize the image from the PTZ camera to the size it would appear in the peripheral view. We then compute the cross-correlation of the resized PTZ image with the peripheral image, and take the actual location of the fovea to be the location with maximum cross-correlation in a small area around its expected location in the peripheral view.

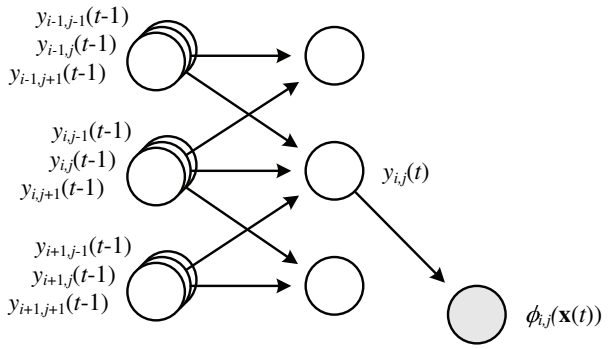

Figure 3: Fragment of dynamic Bayesian network for modeling attentive interest.

The interest belief state over the entire frame at time t is $P(\mathbf{y}(t) \mid \mathbf{y}(0), \mathbf{x}(0), \ldots, \mathbf{x}(t))$. Exact inference in this graphical model is intractable, so we apply approximate inference to estimate this probability. Space constraints preclude a full discussion, but briefly, we applied Assumed Density Filtering/the Boyen-Koller (1998) algorithm, and approximate this distribution using a factored representation over individual pixels, $P(\mathbf{y}(t)) \approx \prod_{i,j} P(y_{i,j}(t))$. In our experiments, this inference algorithm appeared to perform well.

In our model, the interest for a pixel depends both on the interest in the pixel's (motion compensated) neighborhood, $N(i,j)$, at the previous time-step and features extracted from the current frame. The parameterizations for both $P\left(y_{i,j}(t) \mid y_{N(i,j)}(t-1)\right)$ and $P(y_{i,j}(t) \mid \phi_{i,j}(\mathbf{x}(t)))$ are given by logistic functions (with learned parameters), and we use Bayes rule to compute $P(\phi_{i,j}(\mathbf{x}(t)) \mid y_{i,j}(t))$ needed in the DBN belief update. Using this model of the interest map, we estimate the probability of finding an object in a given foveal region, $P(o_k \mid \mathcal{F})$, by computing the mean of the probability of every pixel in the foveal region being interesting.[5]

The features $\phi_{i,j}$ used to predict interest in a pixel are extracted over a local image patch and include Harris corner features, horizontal and vertical edge density, saturation and color values, duration that the pixel has been part of a tracked object, and weighted decay of the number of times the fovea had previously fixated on a region containing the pixel.

3.3 Learning the Interest Model

Recall that we define a pixel to be interesting if it is part of an as yet unclassified object. In order to generate training data so that we can learn the parameters of our interest model, we first hand label *all* classifiable objects in a low-resolution training video sequence. Now as our system begins to recognize and track objects, the pixels associated with those objects are no longer interesting, by our definition. Thus, we generate our training data by annotating as interesting only those pixels marked interesting in our hand labeled training video but that are not part of objects currently being tracked. Using this procedure we can hand label a single training video and automatically generate data for learning the interest model given any specific combination of classifiers and foveal movement policies.

We adapt the parameters of our probabilistic models so as to maximize the likelihood of the training data described above. Our interest model is trained over a 450 frame video sequence, i.e., 30 seconds at 15 frames per second. Figure 4 shows an example of a learned interest map, in which the algorithm automatically selected the mugs as the most interesting region.

Figure 4: Learned interest. The righthand panel shows the probability of interest for each pixel in the peripheral image (left).

4 Experimental Results

We evaluate the performance of the visual attention system by measuring the percentage of times that a classifiable object appearing in the scene is correctly identified. We also count the number of false-positives being tracked per frame. We compare our attention driven method for directing the fovea to three naive approaches:

(i) fixing the foveal gaze to the center of view,

(ii) linearly scanning over the scene from top-left to bottom-right, and,

(iii) randomly moving the fovea around the scene.

Our classifiers are trained on image patches of roughly 100×100 pixels depending on the object class being trained. For each object class we use between 200 and 500 positive training examples and $10,000$ negative examples. Our set of negative examples contained examples of other objects, as well as random images. We extract a subset of C1 features (see [Serre *et al.*, 2004]) from the images and learn a boosted decision tree classifier for each object. This seemed to give good performance (comparable to state-of-the-art systems) for recognizing a variety of objects. The image patch size was chosen for each object so as to achieve good classification accuracy while not being so large so as to prevent us from classifying small objects in the scene.

We conduct two experiments using recorded high-definition video streams, the first assuming perfect classification (i.e., no false-positives and no false-negatives, so that every time the fovea fixates on an object we correctly classify the object), and the second using actual trained state-of-the-art classifiers. In addition to the different fovea control algorithms, we also compare our method against performing object classification on full frames for both low-resolution and

[5] Although one can envisage significantly better estimates, for example using logistic regression to recalibrate these probabilities, the algorithm's performance appeared fairly insensitive to this choice.

high-resolution video streams. The throughput (in terms of frames per second) of the system is inversely proportional to the area scanned by the object classifiers. The low resolution was chosen to exhibit the equivalent computational requirements of the foveal methods and is therefore directly comparable. The high resolution, however, being of much larger size than the foveal region, requires additional computation time to complete each frame classification. Thus, to meet real-time operating requirements, the classifiers were run ten times less often.

Figure 5 shows example timelines for recognizing and tracking a number of different objects using our method (bottom), randomly moving the fovea (middle), and using a fixed fovea (top). Each object is shown on a different line. A dotted line indicates that the object has appeared in the scene but has not yet been recognized (or that our tracking has lost it). A solid line indicates that the object is being tracked by our system. The times when the fovea fixates on each object are marked by a diamond ($\diamond$). It is clear that our method recognizes and starts tracking objects more quickly than the other methods, and does not waste effort re-classifying objects in regions already explored (most apparent when the fovea is fixed).

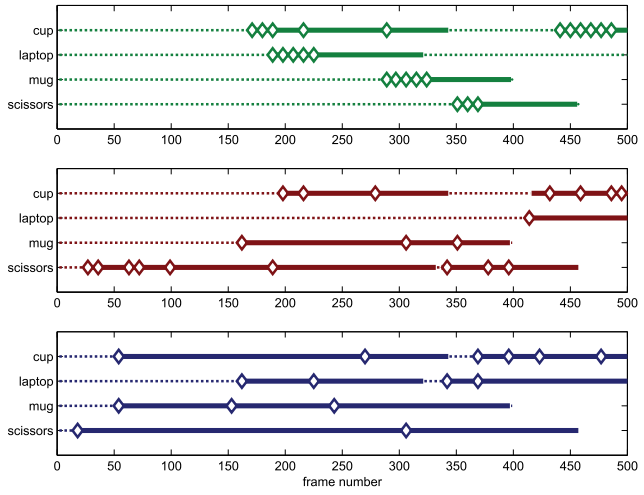

Figure 5: Example timelines for identification and tracking of objects in a continuous video sequence using different methods for controlling the fovea: fixed (top), random (middle), and our interest driven method (bottom).

In our experiments we use a fixed size fovea covering approximately 10% of the peripheral field-of-view.[6] All recognizable objects in our test videos were smaller than the size of the foveal window. We record peripheral and foveal video streams of an office environment with classifiers trained on commonly found objects: coffee mugs, disposable cups, staplers, scissors, etc. We hand label every frame of the video stream. The video sequence consists of a total of 700 frames recorded at 15 frames per second—resulting in 100 foveal fixations (since it takes approximately seven frames for the fovea to move).

A comparison between our method and the other approaches is shown in Table 1. We also include the F_1-score when running state-of-the-art object classifiers in Table 2.

Fovea control	Perfect classification	Actual classifiers
Full image (low-resolution)	n/a	0.0%[a]
Full image (high-resolution)	n/a	62.2%
Fixed at center	16.1%	15.9%
Linear scanning	61.9%	43.1%
Random scanning	62.0%	60.3%
Our (attentive) method	**85.1%**	**82.9%**

[a] Our object classifiers are trained on large images samples (of approximately 100×100 pixels). This result illustrates that objects in the low-resolution view occupy too few pixels for reliable classification.

Table 1: Percentage of objects in a frame that are correctly identified using different fovea control algorithms.

Fovea control	Recall	Precision	F_1-Score
Full image (low-res)	0.0%	0.0%	0.0%
Full image (high-res)	62.2%	95.0%	75.2%
Fixed at center	15.9%	100.0%	27.4%
Linear scanning	43.1%	97.2%	59.7%
Random scanning	60.3%	98.9%	74.9%
Our method	82.9%	99.0%	**90.3%**

Table 2: Recall, precision and F_1-score for objects appearing the scene. The low-resolution run results in an F_1-score of zero since objects appearing in the scene occupy too few pixels for reliable classification.

The results show that our attention driven method performs significantly better than the other foveal control strategies. Our method even performs better than scanning the entire high-resolution image. This is because our method runs much faster since it only needs to classify objects in a small region. When running over the entire high-resolution image, the system is forced to skip frames so that it can keep up with real-time. It is therefore less able to detect objects entering and leaving the scene than our method. Furthermore, because we only direct attention to interesting regions of the scene, our method results in significantly higher precision, and hence better F_1-score, than scanning the whole image.

Finally, we conduct experiments with a dual wide-angle and PTZ camera system. In order to compare results between the different foveal control strategies, we fix the robot and scene and run system for 100 frames (approximately 15 foveal fixations) for each strategy. We then change the scene by moving both the objects and the robot pose, and repeat the experiment, averaging our results across the different scenes. The results are shown in Table 3. Again our attention driven method performs better than the other approaches.

Videos demonstrating our results are provided at

http://ai.stanford.edu/∼sgould/vision/

[6] Although direct comparison is not possible, for interest we note that the human fovea covers roughly 0.05% of the visual field.

Fovea control	Recall	Precision	F_1-Score
Fixed at center[a]	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%**

[a] In some of the trials a single object happened to appear in the center of the field of view and hence was detected by the stationary foveal gaze.

Table 3: Recall, precision and F_1-score of recognizable objects for hardware executing different foveal control strategies over a number of stationary scenes.

5 Discussion

In this paper, we have presented a novel method for improving performance of visual perception on robotic platforms and in digital video. Our method, motivated from biological vision systems, minimizes the uncertainty in the location of objects in the environment using a with controllable pan-tilt-zoom camera and fixed wide-angle camera. The pan-tilt-zoom camera captures high-resolution images for improved classification. Our attention system controls the gaze of the pan-tilt-zoom camera by extracting interesting regions (learned as locations where we have a high probability of finding recognizable objects) from a fixed-gaze low-resolution peripheral view of the same scene.

Our method also works on a single high-resolution digital video stream, and is significantly faster than naively scanning the entire high-resolution image. By experimentally comparing our method to other approaches for both high-definition video and on real hardware, we showed that our method fixates on and identifies objects in the scene faster than all of the other methods tried. In the case of digital video, we even perform better than the brute-force approach of scanning the entire frame, because the increased computational cost of doing so results in missed frames when run in real-time.

Acknowledgments

We give warm thanks to Morgan Quigley for help with the STAIR hardware, and to Pieter Abbeel for helpful discussions. This work was supported by DARPA under contract number FA8750-05-2-0249.

References

[Björkman and Kragic, 2004] Marten Björkman and Danica Kragic. Combination of foveal amd peripheral vision for object recognition and pose estimation. In *Proceedings of the 2004 IEEE International Conference on Robotics and Automation*, pages 5135–5140, 2004.

[Boyen and Koller, 1998] Xavier Boyen and Daphne Koller. Tractable inference for complex stochastic processes. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence—UAI 1998*, pages 33–42. San Francisco: Morgan Kaufmann, 1998.

[Brubaker *et al.*, 2005] S. Charles Brubaker, Jianxin Wu, Jie Sun, Matthew D. Mullin, and James M. Rehg. On the design of cascades of boosted ensembles for face detection. In *Tech report GIT-GVU-05-28*, 2005.

[Edelman and Weiss, 1995] Shimon Edelman and Yair Weiss. Vision, hyperacuity, 1995.

[Fahle and Poggio, 1981] M. Fahle and T. Poggio. Visual hyperacuity: Spatiotemporal interpolation in human vision. In *Proceedings of the Royal Society of London*. The Royal Society, 1981.

[Hu *et al.*, 2004] Yiqun Hu, Xing Xie, Wei-Ying Ma, Liang-Tien Chia, and Deepu Rajan. Salient region detection using weighted feature maps based on the human visual attention model. In *Fifth IEEE Pacifc-Rim Conference on Multimedia*, Tokyo Waterfront City, Japan, November 2004.

[Itti and Koch, 2001] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, Mar 2001.

[Itti *et al.*, 1998] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[Klein, 2000] Raymond M. Klein. Inhibition of return. *Trends in Cognitive Sciences*, 4(4), April 2000.

[Koch and Ullman, 1985] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[Lucas and Kanade, 1981] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision (ijcai). In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81)*, pages 674–679, April 1981.

[Orabona *et al.*, 2005] Francesco Orabona, Giorgio Metta, and Giulio Sandini. Object-based visual attention: a model for a behaving robot. *IEEE Conference on Computer Vision and Pattern Recognition*, 3:89, 2005.

[Privitera and Stark, 2000] Claudio M. Privitera and Lawrence W. Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):970–982, 2000.

[Serre *et al.*, 2004] Thomas Serre, Lior Wolf, and Tomaso Poggio. A new biologically motivated framework for robust object recognition. In *Al Memo 2004-026*, November 2004.

[Tagare *et al.*, 2001] Hermant D. Tagare, Kentaro Toyama, and Jonathan G. Wang. A maximum-likelihood strategy for directing attention during visual search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(5):490–500, 2001.

[Ude *et al.*, 2003] A. Ude, C. G. Atkeson, and G. Cheng. Combining peripheral and foveal humanoid vision to detect, pursue, recognize and act. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, pages 2173–2178, October 2003.

[Ude *et al.*, 2006] A. Ude, C. Gaskett, and G. Cheng. Foveated vision systems with two cameras per eye. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pages 3457–3462, May 2006.

[Viola and Jones, 2004] Paul Viola and Michael J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[Wandell, 1995] Brian A. Wandell. *Foundations of Vision*. Sinauer Associates, Inc., 1995.

[Zhang, 2000] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.