2008

Combined object recognition approaches for mobile robotics

Rusty Gerard
Western Washington University

Combined Object Recognition Approaches for Mobile Robotics

Rusty Gerard
gerardr@cc.wwu.edu
Computer Science Department
Western Washington University

Abstract

There are numerous solutions to simple object recognition problems when the machine is operating under strict environmental conditions (such as lighting). Object recognition in real-world environments poses greater difficulty however. Ideally mobile robots will function in real-world environments without the aid of fiduciary identifiers. More robust methods are therefore needed to perform object recognition reliably. A combined approach of multiple techniques improves recognition results.

Active vision and peripheral-foveal vision—systems that are designed to improve the information gathered for the purposes of object recognition—are examined. In addition to active vision and peripheral-foveal vision, five object recognition methods that either make use of some form of active vision or could leverage active vision and/or peripheral-foveal vision systems are also investigated: affine-invariant image patches, perceptual organization, 3D morphable models (3DMMs), active viewpoint, and adaptive color segmentation.

The current state-of-the-art in these areas of vision research and observations on areas of future research are presented. Examples of state-of-the-art methods employed in other vision applications that have not been used for object recognition are also mentioned. Lastly, the future direction of the research field is hypothesized.

Keywords: computer vision, object recognition, mobile robotics, active vision, peripheral-foveal vision.

1: Introduction

Object recognition, especially for mobile robotics, is a challenging area of computer vision. In the case of machine vision and industrial robotics [1] there are numerous solutions to simple object recognition problems when the machine is operating under strict conditions of lighting, orientation, occlusion and distractors. Because mobile robots are typically expected to operate in the real-world and not a sterilized environment, they seldom have such luxuries and thus more robust solutions are needed. A combined approach of multiple techniques can be employed to improve results [2]-[5] [7] [9].

1.1: Overview of Principle Methods of Discussion

This paper is primarily concerned with active vision systems and peripheral-foveal vision systems. Briefly, these two types of vision systems are designed to improve the information gathered for the purposes of object recognition [2]-[5]. In addition to active vision and peripheral-foveal vision, five object recognition methods that either make use of some form of active vision or could leverage active vision and/or peripheral-foveal vision systems are also investigated: affine-invariant image patches [6], perceptual organization [7], 3D morphable models (3DMMs) [8], active viewpoint [9], and adaptive color segmentation [10].

1.2: Organization of this Paper

The rest of this paper is organized as follows: Section 2 discusses each method in detail, Section 3 reports the experimental results of those methods, Section 4 discusses the key points—i.e. advantages and disadvantages—of the presented research, Section 5 presents areas of inspiration and future work in these areas of research, and lastly Section 6 is concluding remarks.

1.3.1: Table of Contents

2: Methods

Section 2 describes the principal methods of discussion, the current state-of-the-art in those areas and specific algorithms. Each of the seven subsections describes one method, though due to the nature of the research in this field there is a degree of overlap and intermingling between those methods.

2.1: Active Vision

An active vision (aka active computer vision) system investigates its environment by moving the camera system in order to gain more information, in this case with a robot-mounted camera. For example, when viewing a point from multiple angles, features in the background tend to move or change more drastically than features in the foreground do. Rather than only passively examining images of a scene in an attempt to recognize objects, active vision can be employed in addition to other object recognition methods. Doing so enables the system to more readily separate objects in the foreground from objects in the background [2], and circumvent occlusion [2] [3].

Kootstra, Ypma, and de Boer [2] have implemented an active vision system using Lowe's Scale-Invariant Feature Transform (SIFT) keypoints and an optimization approach to next viewpoint selection. To reduce the computational complexity of SIFT they have used an alternative tracking system based on the Marsland-Shapiro-Nehmzow Growing When Required (GWR) neural network. The computational complexity is reduced thanks to the GWR network's fewer necessary keypoints for successful recognition. The tradeoff is reduced effectiveness as the network grows—an unavoidable fact as the network is tasked with recognizing multiple objects.

Farshidi, Sirouspour, and Kirubarajan [3] have experimented with an active-vision system with the intent of maximizing the object recognition success rate while keeping the necessary number of images collected to a minimum using a Bayesian probabilistic approach. The probabilistic framework models camera noise, occlusions, and errors in the camera and object positioning systems. The available information is used to determine the next position of the cameras so that the most information about the object can be obtained. Operating under the assumption that the appearance of objects to be recognized will have a high degree of similarity from different viewpoints, Farshidi, Sirouspour and Kirubarajan implemented an appearance-based parametric eigenspace method. Training images are converted to vectors in a matrix; a small number of eigen-images corresponding to the highest eigenvalues are saved.

One of they key areas of study with active vision is the maximization of information gathered with the minimum number of images. Note that in Kootstra, Ypma, and de Boer's implementation above [2], 36 images are captured for the initial information step, and then new viewpoints are selected to gather additional information. Farshidi, Sirouspour, and Kirubarajan [3] have published preliminary results showing their active vision system has a high object recognition and pose estimation success rate (often higher than 95%) using no more than seven pairs of images.

A number of other researchers included later in the paper have made use of active vision to some extent in conjunction with their principal research. For example, in their experiments with perceptual organization, Shang, Ma and Dai [7] employ an active vision step in their algorithm to center the target object in the camera's frame of view. Forssén et al. [5] employ an active vision step for the purposes of viewpoint gathering in order to gain novel views of objects. For more information refer to the following sections related to peripheral-foveal vision (section 2.2, pp. 6-8) and perceptual organization (section 2.4, pg. 10).

2.1.1: Kootstra-Ypma-de Boer Algorithm

Below is Kootstra, Ypma, and de Boer's [2] active vision and keypoint clustering algorithm. The algorithm is designed to simplify recognition tasks by: separating the object from its background using what they call a, "what-moves-together-belongs-together" approach; detecting stable keypoints; and novel viewpoint collection.

1. Active vision
    a. Search for stable keypoints:
        i. Circle the object capturing images 10° apart
        ii. Match the keypoint to its nearest neighbor in the adjacent image:
        iii. For each keypoint in the first image:
            For each keypoint in the second image:
                Compare the appearances of the keypoints to match them
        iv. For each keypoint pair:
            If the Euclidian distance between the paired keypoint feature vectors is large, the keypoints are unstable
    b. Segment background keypoints from foreground keypoints:
        For each keypoint pair:
            If the horizontal keypoint disparity $|x_1 - x_2|$ is less than the horizontal threshold and the vertical keypoint disparity $|y_1 - y_2|$ is less than the vertical threshold the keypoints are part of the foreground
    c. Determine confidence of object recognition:
        For all 36 poses (images):
            For every stable foreground keypoint in the image:
                For every known object:
                    If the visual description of the keypoint is associated with the object, increase the confidence of the keypoint/object/pose group
2. Next viewpoint selection
    Determine the pose (angle) that maximizes the expected confidence:
        Optimize the expected confidence function over the keypoint/object/pose group
3. Keypoint clustering
    (Modified Marsland-Shapiro-Nehmzow GWR neural network)
    Cluster similar keypoints:
        For each keypoint:
            a. Input the keypoint/object/pose group to the network
            b. Select the best matching (winning) node
            c. Calculate the activity of the winning node
            d. Calculate the firing counter of the winning node
            e. If the keypoint is sufficiently different from the winning node and the firing counter is below a threshold, add a new node to the network
            f. If the keypoint is sufficiently similar to the winning node, cluster it with that node

2.1.2: Farshidi-Sirouspour-Kirubarajan Algorithm

Below is Farshidi, Sirouspour, and Kirubarajan's [3] active multi-camera algorithm. The algorithm is designed to minimize the number of images captured while maximizing the information about the object. Images are converted to eigenspaces, and then the eigenvectors that correspond to the largest eigenvalues are collected and used to represent the object. The key contributions of this algorithm are the formulation of a multi-camera active recognition system and the modeling of the occlusion level in the Bayesian network.

1. Model each state in the Bayesian network
   Estimate pdf of unknown state n based on known observations and sensor parameters and the estimated pdf of state n – 1:
   ```
   For each object class:
       For each pose:
           For each occlusion level in camera 1:
               For each occlusion level in camera 2:
                   Accumulate probability density of state n
   ```
2. Select next sensor location
   Set the position of the next state to be the argument of the maximum mutual information (the state that minimizes information uncertainty and ambiguity)

2.2: Peripheral-Foveal Vision

The human retina can be divided into two components: the periphery which consists mainly of light-sensitive photoreceptor cells (aka rods), and the fovea which consists mainly of color-sensitive photoreceptor cells (aka cones). The central fovea is primarily responsible for detailed vision, while the surrounding periphery of the retina is primarily responsible for detecting motion and observing the non-central field of view [4]. The concept of peripheral-foveal vision is to emulate these aspects of human vision: a low-resolution wide-angle image is first used to detect points of interest in the scene, and then those points are further investigated using conventional vision techniques using a high-resolution image. This technique can be implemented using either a low-resolution wide angle camera paired with a high-resolution pan-tilt-zoom camera [4] [5] or using multiple resolutions of the same image captured by a single high-resolution wide angle camera [4].

Gould et al.'s [4] system uses what is called a "Visual Attention Model". In short, the low resolution wide-angle camera is used to select features and generate an "attentive interest map". Based on the interest map the high resolution camera is centered on an area to either search for unidentified objects in the scene or to confirm the positions of objects in the scene that have been tracked over time. The interest map is calculated by labeling each pixel with a level of interest, the level of interest being a combination of the probability that the pixel belongs to an object that the system is trained to recognize and that that object is unknown. Each pixel is modeled over time using a Dynamic Bayesian Network (DBN). Determining the exact probability for each pixel in the DBN is intractable, so the probability is instead approximated using Assumed Density Filtering (aka the Boyen-Koller algorithm). Stereo calibration is used to transfer objects of interest in the foveal camera's coordinate space to the peripheral camera's coordinate space so that they may be tracked over time. Objects are tracked over time by the low-resolution camera using a Kalman filter.

Forssén et al. [5] attempts to solve the lost-and-found problem, i.e. locate an object or set of objects out of a larger set of objects. To do this, they have implemented a peripheral-foveal system: the low-resolution peripheral cameras scan the area for regions of interest that may correspond to objects in the world; those interesting regions are then examined with the high-resolution foveal camera. The difference between this system and previous work, including the paper on peripheral-foveal vision by Gould et al. [4] above, is that those systems track previously recognized objects, and depend on reliable recognition, in the peripheral images. The robot maps its environment using a fast simultaneous localization and mapping (SLAM) algorithm, examining objects as it does so and noting the locations of those objects so that they may be re-examined from alternate angles later (i.e. the informed visual search). The behavior of the robot can be summarized as three stages: 1) exploration (mapping unknown areas with a range sensor), 2) coverage (examining the mapped environment with the peripheral cameras), and 3) viewpoint selection (acquiring additional images of interesting objects from varying angles with the foveal camera).

It should be noted that Forssén et al.'s approach [5] utilizes elements of active vision, specifically the multiple viewpoint collection stage of gaze planning. For more information, refer to the previous section (active vision, section 2.1, pp. 4-6) and the overview of Forssén et al.'s algorithm below (section 2.2.2, pg. 8).

2.2.1: Gould et al.'s Algorithm

Below is Gould et al.'s [4] peripheral-foveal vision algorithm. The algorithm is designed to effectively detect areas of interest in streaming video (i.e. the peripheral view), identify objects in the foveal view, and detected objects over time in the peripheral view robustly without dropped frames.

1. Visual attention
   Periodically decide which action to perform:
   a. Object position confirmation:
      Fixate the fovea on a predicted location of an object to confirm its presence and update its position estimate
   b. Object search:
      Fixate the fovea on a new area and run the object classifier
2. Object tracking
   For each image frame:
   For each tracked object:
   a. Compute object velocity
      (Lucas-Kanade sparse optical flow algorithm)
   b. Kalman filter motion update using position and velocity estimates
   c. Kalman filter observation update
3. Generate interest model
   Estimate probability that unknown regions contain unknown, identifiable (interesting) objects:
   For each pixel in the peripheral:
   Determine probability pixel belongs to an interesting object:
   a. Model the pixel features (used by the object classifier) in a Dynamic Bayesian Network (DBN)
   b. Estimate probability that pixel is interesting (unknown and identifiable)
      (Boyen-Koller Assumed Density Filtering algorithm)

2.2.2: Forssén et al.'s Algorithm

Below is Forssén et al.'s [5] informed visual search algorithm. The algorithm is designed to solve the lost-and-found problem (finding a set of objects that are known to be present in the environment). The system combines active vision and peripheral-foveal vision techniques to more effectively identify objects using a bag-of-features approach. The peripheral-foveal system detects and identifies objects while the active vision system plans paths to collect novel information about as many detected objects in the environment as possible rather than exploring each object in turn.

1. Initial Exploration
   Produce geometric map of nearby traversable regions
   (Yamauchi-Schultz-Adams Frontier Map-building and Localization algorithm)
2. Visual Attention
   Select potential objects:
   (Hou-Zhang Spectral Risidual Saliency Detector)
   a. Compute saliency on intensity, red-green, and yellow-blue channels
   b. Combine saliency maps
   c. Segment regions in the combined map
      (Maximally Stable Extremal Region Detector)
3. Gaze Planning
   a. Multiple Viewpoint Collection
      Select a goal location which will provide new information for the most objects:
      For each object:
      i. Select the best location for viewing the object
      ii. If the location is reachable by the robot and the pose to be collected will be new
          Vote for that location
   b. Center a potential object in the foveal camera
      (Forssén Saccadic Gaze Control Algorithm)
4. Continuous Geometric Mapping and Navigation
   (Montemerlo et al. FastSLAM 2.0 algorithm)
   (Borenstein et al. Vector Field Histogram algorithm)
5. Object Recognition
   a. Bag-of-features method
      i. For each feature in the observed image
         For each feature in the training images
            Match features using nearest neighbor approach
      ii. For all matched features:
         - Compute ratio of distances to foreground and background
         - If the ratio is high, discard the feature
         - Compute bag-of-features score
   b. Geometric matching
      i. For all matched features in the observed image
         For all matched features in the training image
            Compute similarity transform
      ii. Compute goodness-of-fit of the similarity transform
      iii. If the fit is good, the observed image contains a known object

2.3: Affine-Invariant Image Patches

Affine invariants (i.e. a viewpoint that does not change regardless of the orientation of an object) are, generally speaking, only effective at identifying planar objects and simple shapes and are not effective at recognizing 3D objects [6]. Rather than using invariants to describe an entire 3D object, Rothganger et al. [6] propose using invariants to describe small planar patches on the surface of 3D objects. These invariant image patches can then be used to index objects for the purpose of recognition. The object is represented by the appearance of the patches, the invariants of the patches, and the spatial relationship of the patches.

2.3.1: Rothganger et al.'s algorithm:

Below is Rothganger et al.'s [6] affine-invariant image patch indexing algorithm. The algorithm is designed to represent 3D objects as a set of small, affine-invariant patches and the spatial relationship of those patches.

1. Create indexing patches:
   For each view (image) of the object:
   a. Determine points of interest (POI):
      (Mikolajczyk-Schmid affine-invariant region detector)
      For each image patch:
      i. Normalize the shape of the patch
      ii. Determine the characteristic scale of the patch's brightness pattern
      iii. Determine the patch's location
      iv. Eliminate rotational ambiguity
   b. Match points of interest from the current image with points from other images:
      For each resultant POI (normalized patch and affine transformation matrix):
      For each adjacent image:
      i. Select the N most similar POIs based on appearance and affine transform
      ii. Find groups of consistent matches:
      For each match:
      1. Find the match that minimizes the inconsistency of the group
      2. If the minimized match passes a threshold test, add it to the group
      iii. Discard all groups that are too small
2. Construct the model:
   a. Split the data into overlapping blocks of multiple images
   b. For all patches in all images of the same block
      i. Use the matching strategy from 1b
   c. For all points common to overlapping blocks
      i. Register the reconstructions
      ii. Initialize variable estimates
   d. Refine variable estimates

2.4: Perceptual Organization

According to Shang, Ma, and Dai, [7] perceptual organization—the description of objects based upon Gestalt principles such as symmetry and similarity—is a powerful tool for object recognition. Shang, Ma, and Dai have developed a perceptual organization recognition system based on Bayesian networks. The system describes 3D structural objects, like doors and windows and other objects that are common in office environments, in terms of lines and their spatial relationship to each other. A door for example can be modeled in a Bayesian network as two vertical line segments connected by a horizontal line segment at the top. Other networks can model windows, cubicles, etc. The extracted features are handed off to each Bayesian network and each network is executed in parallel. To test their methods, Shang, Ma, and Dai implemented a cubicle-recognition network and a door-recognition network.

2.4.1: Shang-Ma-Dai Algorithm:

Below is Shang, Ma, and Dai's [7] perceptual organization algorithm. The algorithm recognizes objects by modeling in a Bayesian network each object to be recognized by line segments and the spatial relationship of those line segments. An active vision step is employed to center the object of interest in the camera's field of view.

1. Low-level feature extraction:
    a. Search for straight lines
    b. Remove gaps in lines:
        i. Search for breaks in lines
        ii. If a line parallel to the gap exists and the length is less than a threshold value, connect the broken line
2. Parallel Bayesian network execution:
    a. Robot movement step:
        i. Determine distance to observed object using range sensors
        ii. If distance from observed object to robot is not between min. and max., move the robot
        iii. If features are extracted that approach the image borders, adjust the camera viewing angle
        iv. If the camera viewing angle has been adjusted to the maximum, move the robot forward or back
    b. For each network:
        i. For each evidence node in the network:
            Determine if the evidence node is true or false based on the extracted low-level features
        ii. Based on the known evidence, compute the conditional probability that the extracted features describe the polyhedron described by the network
        iii. If the probability is greater than a threshold, the polyhedron described by the network has been detected

2.5: 3D Morphable Models

According to Romdhani and Vetter [8], 3D morphable models (3DMM) are considered to be the state of the art in object recognition, particularly face recognition. 3DMMs are highly tolerant to combined variations in pose and illumination in the image and require only a few feature points for recognition. Unfortunately, with most implementations, these feature points need to be selected by hand with each image that is to be analyzed and there is no known 3DMM implementation that is fully automated. Romdhani and Vetter have developed an automated system for the selection of feature points for a recognition system based upon 3D Morphable Face Models (3DMFMs).

Romdhani and Vetter's [8] feature point model is viewpoint invariant and shares some similarities to Rothganger et al.'s viewpoint invariant image patch method above. The key difference between the two methods is that Romdhani and Vetter have constructed a generic model specifically for faces and a single image is used to fit an object to the model as best as possible—the parameters used to fit the specific image to the generic model are used for recognition purposes, whereas Rothganger et al.'s method [6] captures multiple views of an object and constructs a model for identifying that particular object.

2.5.1: Romdhani-Vetter Algorithm:

Below is Romdhani and Vetter's [8] feature point selection algorithm. The algorithm is designed to automatically detect key feature points for a 3D object model based on 3DMFMs.

1. SIFT point detection:
   (Lowe's SIFT Detector)
   a. Gather a set of key point locations along with their scale and orientation from the image
   b. For each key point:
      Construct a normalized image patch around the key point
2. Appearance model rejection:
   For each key point:
      For each feature point set
         a. Calculate the appearance likelihood ratio
         b. If the ratio is not small, add the key point to the feature point set
         c. Keep only the top N key points
3. Projection constraint rejection:
   a. For each reference point:
      For each feature point set:
         i. Construct the hypothetical 3D to 2D projection matrix
         ii. If the matrix does not conform to the geometric constraints, reject it
   b. For each valid matrix:
      Extract the projection parameters
4. Maximum likelihood estimate:
   a. Minimize the log likelihood ratio:
      For each reference point:
         For each (remaining) feature point set:
            Compute the likelihood ratio
   b. If the likelihood ratio is large, no face is detected in the image
5. Parameter refinement:
   If a face was detected:
      Maximize the maximum likelihood ratio:
      (Gauss-Newton Algorithm similar to Fischer-Bolles RANSAC)

2.6: Active Viewpoint

Yabuta, Mizumoto, and Arii [9] propose a novel approach to efficient shape recognition using a combination of a stereo camera pair and a pair of lasers that they dub “active viewpoint”. When the cameras fix on an object the lasers intersect at the point of focus. This point of intersection is used by the system to center both cameras’ viewpoints on the same target, eliminating the need the need for stereo matching computation so that triangulation and depth estimation can be computed directly. The system is suited for target objects that have a few distinct feature points along the projected line of correspondence (such as polyhedrons), so the authors have proved their system by applying it to recognizing cubes in various poses.

To recognize cubes, edge data is extracted from the image and used for Hough transformations. Regions surrounded by four straight lines are isolated to compute quadrangles (rhombus, rectangle, square, etc.) and the aspect ratios of those quadrangles. These quadrangles and their aspect ratios in both the left and right images captured by the stereo pair are used to compute the spatial coordinates of the object.

2.6.1: Yabuta-Mizumoto-Arii Algorithm:

Below is Yabuta, Mizumoto, and Arii’s [9] active viewpoint algorithm. The algorithm is designed to reduce the computational complexity of stereo correspondence for the purposes of object recognition. Using a pair of lasers in conjunction with the cameras, lines of correspondence are projected and searched along rather than searching the entire image. The system is demonstrated by estimating the pose and size of cubes.

1. Determine camera viewpoint:
    a. The point targeted by the lasers is the fixation point
    b. Center the projected point in both the left and right images
2. Calculate distance from stereo camera pair to fixation point:
    a. Determine the correspondence from candidate features along the right epipolar line
    b. Eliminate false features using the active viewpoint
    c. Extract the feature point from the left image
    d. Triangulate
3. Extract object surfaces:
   (Color-segment the image)
4. Detect edges
5. Detect straight lines:
   (Hough transform)
6. Identify quadrangles
7. Construct right-left image correspondence
8. Calculate object’s spatial coordinates

2.7: Adaptive Color Segmentaiton

Color-based object recognition is a popular method for fast object recognition in several applications of mobile robotics, particularly robot soccer [10]. Objects of interest are uniquely color-coded and the color-coding scheme is made known to the robot. This type of object recognition is highly susceptible to variations in lighting, particularly when outdoors however. Browning and Veloso [10] have developed a novel solution to the variable lighting problem that leverages the fact that colors change in a continuous way under variable lighting conditions. In other words, the color red will always appear to be a shade of red regardless of the lighting conditions. By using an adaptive threshold, Browning and Veloso are able to associate many shades of a color to one class of object(s).

2.7.1: Browning-Veloso Algorithm:

Below is Browning and Veloso's [8] adaptive color segmentation algorithm. The algorithm is designed to robustly segment the image by color regardless of lighting and shading conditions. This is accomplished by using an adaptive threshold technique where the color class threshold that fully encompasses the color class histogram is used for segmentation.

1. Segment the image:
   For each pixel in the image:
      For each object color class:
         Assign pixel to the color class it is most likely to belong to
2. Build histograms:
   (Fu-Gonzales-Lee Color Histogram Algorithm)
3. Adapt thresholds:
   For each object color class:
   a. Convolve the histogram with a zero-mean Gaussian kernel
   b. Calculate the stationary points of the convolved histogram
   c. Label each stationary point as inflection, peak, or trough
   d. Set the target threshold to the argument of the maximum trough (the trough that most fully captures the peak and its sides)
   e. If the target threshold is greater than a fixed minimum value (determined *a priori*), move the actual threshold partway towards the target threshold
4. Find objects:
   a. Extract regions of similarly labeled pixels using component analysis:
      (CMVision Length Encoding and Conglomeration Algorithm)
   b. Object Recognition:
      (Lenser-Bruce-Veloso Object Recognition Algorithm)
      For each region:
      i. Eliminate unrecognizable regions based on functions of: area, bounding box dimensions, bounding box density, and expected size and area based its location in the image or relation to neighboring regions
      ii. Model each region with a Gaussian uncertainty model (mean and variance determined *a priori*)
      iii. Determine likelihood that region is one of possible interest
      iv. Eliminate regions of low likelihood

3: Experimental Results

Sections 3.1 and 3.2 below briefly summarize the experimental results of the methods outlined in this paper. Results in each table are organized by principal method of discussion, then by author.

3.1: Summary of Experimental Results

Table 1 below summarizes the recognition rate, recall and precision (when available) of the methods outlined in this paper. Table 1 is not intended as basis for comparison between algorithms or methods; it is merely a summary of experimental results. Every paper uses different data sets for testing and presents their results in different ways. Even when results are presented using the same metrics the fact that different test data and hardware was used between experiments makes direct comparison difficult.

Table 1: Summary of Experimental Results

Method	Authors	Recognition Rate	Recall	Precision
Active Vision	Kootstra et al.	90%	Not specified	Not specified
	Farshidi, Sirouspour, and Kirubarajan	96%	Not specified	Not specified
Peripheral-Foveal Vision	Gould et al.	Not specified	62%	83%
	Forssén et al.	75%	Not specified	Not specified
Affine-Invariant Image Patches	Rothganger et al.	Not specified	Not specified	Not specified
Perceptual Organization	Shang, Ma, and Dai	Not specified	Not specified	Not specified
3DMMs	Romdhani and Vetter	92%	Not specified	Not specified
Active Viewpoint	Yabuta, Mizumoto, and Arii	Not specified	Not specified	Not specified
Color Segmentation	Browning and Veloso	Not specified	Not specified	Not specified

3.2: Comparison of Methods

Table 2 below summarizes the key features of the experimental results for each of the methods outlined in this paper. The information provided is meant to give insight into the relevance of the data provided in Table 1 and the algorithms outlined in Section 2 above.

Table 2: Comparison of Key Features

Method	Authors	Key Features of the Experimental Results
Active Vision	Kootstra et al.	Performance diminishes as the object database increases.
	Farshidi, Sirouspour, and Kirubarajan	Experiments are conducted with "perfect knowledge" e.g. perfect camera zoom levels are known a priori for every camera position.
Peripheral-Foveal Vision	Gould et al.	Training sets consist of 200-500 positive examples and 10,000 negative examples.
	Forssén et al.	System tested using 4 uniformly-spaced training images of each object to be recognized.
Affine-Invariant Image Patches	Rothganger et al.	Reprojection error is consistently small (i.e. recognizing the model that corresponds to the object of interest and placing the model in the correct location and pose), even with large, high-resolution images.
Perceptual Organization	Shang, Ma, and Dai	Recognition results are consistently reliable regardless of robot position, object size/shape, occlusion and distractors commonly found in an office environment.
3DMMs	Romdhani and Vetter	Training set consisted of 871 front and side images from the FERET image database. Experiments were conducted using a subset of the CMU-PIE image database consisting of 68 individuals.
Active Viewpoint	Yabuta, Mizumoto, and Arii	Experiments were conducted on cubes. Each visible face of the cube had a unique color (white/yellow/red) to aid in the quadrangle detection step.
Color Segmentation	Browning and Veloso	Image segmentation is reliable under varying lighting and shade conditions in the image, provided that the color lookup tables are calibrated correctly. E.g. results on a cloudy day are similar to the results on a sunny day.

4: Discussion

As a follow-up to section 3 above, Table 3 below outlines the key advantages and disadvantages of the presented research. In particular, details that might not be apparent from sections 2-3 above or that are also noteworthy but were not included in those sections are outlined.

Table 3: Summary of Key Advantages and Disadvantages

Method	Authors	Advantages	Disadvantages
Active Vision	Kootstra et al.	Effectively segments the foreground object from background distractors. Performs better than standard SIFT approaches when using fewer keypoints.	Performs worse than standard SIFT approaches when using more keypoints. The GWR network has scalability issues.
	Farshidi, Sirouspour, and Kirubarajan	Minimizes number of images to be captured for recognition/pose estimation. Functions effectively when occlusion is present.	Has not been tested in the "real world", i.e. parameters such as pre-determined optimal camera zoom must be relaxed.
Peripheral-Foveal Vision	Gould et al.	Attention system yields precision and recall on par with linear scan techniques and is fast enough for real-time video without dropped frames.	Recognition is not possible when objects are either too small to be tracked by the peripheral camera or too large to be viewed entirely by the foveal camera.
	Forssén et al.	Combines a peripheral-foveal vision system with an active vision system.	Does not fully exploit active vision potential, e.g. vertical view planning like with Farshidi et al.'s [3] system.
Affine-Invariant Image Patches	Rothganger et al.	Capable of generating a viewpoint-invariant model for virtually any 3D object.	Does not extend to the perspective case (i.e. when the camera-object distance changes more than slightly).
Perceptual Organization	Shang, Ma, and Dai	Robust to changes in scale, color, lighting conditions, distractors and occlusion.	A Bayesian network must be designed by hand for each object to be recognized.
3DMMs	Romdhani and Vetter	Has a high recognition success rate and is viewpoint-invariant.	Algorithm is specific to shape models such as 3DMMs and 3DMFMs.
Active Viewpoint	Yabuta, Mizumoto, and Arii	Efficient and effective when the target object has a few distinct features along the projected line of correspondence	Has not been proven on complicated 3D objects.
Color Segmentation	Browning and Veloso	Fast. (Mean < 9ms, standard deviation < 1ms.)	Objects of interest must be uniquely color-coded.

5: Inspiration and Direction of Future Research

The success of an active vision algorithm lies in its ability to initially identify an object of interest so that the object of interest can be explored. Likewise, the success of a peripheral-foveal algorithm is in its ability to effectively identify areas of interest for detailed study. It is clear that these two vision systems complement each other as is evident by their joint usage in systems such as [4] and [5]. Typically a single low-resolution, wide-angle camera is employed for the peripheral view [4] [5], though Forssén et al. [5] experimented with a stereo camera pair as the peripheral vision system. These researchers have concentrated on vision using conventional camera systems and recognition using conventional identification techniques.

5.1: Areas of Exploration

Sections 5.1.1—5.1.4 below describe areas of research that either have not yet been explored with respect to active vision and/or peripheral-foveal vision, or need to be explored in greater depth.

5.1.1: Increase the Field-of-View

For the purposes of object detection and tracking, a narrow-angle camera can be detrimental; hence wide-angle cameras are preferred [5]. However, a conventional wide-angle camera does not leverage the full potential of such a vision system. An omni-directional peripheral camera system would enable a robot to track and/or detect objects of interest regardless of their relative position. This could be accomplished by using several cameras positioned around the robot with panoramic stitching or by using a catadioptric camera system.

Catadioptric camera systems using vertically-aligned hyperboloidal mirrors are able to generate panoramic or stereo panoramic images and have been employed with success for single-camera simultaneous localization and mapping [12] [13]. The omnidirectional vision system was superior to conventional cameras for feature detection, feature tracking and outlier rejection in cases where moving objects (i.e. people) partially occluded the scene and where areas of the environment were featureless, but it does not make use of active vision or peripheral-foveal [13]. To date we know of no application of such a system to active vision or peripheral-foveal vision techniques.

5.1.2: Increase the Number of Perspectives

As was mentioned previously, Forssén et al. experimented with a peripheral-foveal system that used a binocular peripheral view. It has been long been established that the benefits outweigh the costs of trinocular systems over their binocular counterparts, namely reduced disparity errors compared to increased computational expense [14]. A trinocular system could likewise be employed for the peripheral view, or a binocular system could be employed for the foveal view of a peripheral-foveal system. The use of a binocular or trinocular foveal camera system is particularly intriguing as depth-based recognition techniques could be combined with the peripheral-foveal system. This approach could be further generalized to polynocular (n-camera) peripheral-foveal systems.

5.1.3: Alternative Interest Detection Techniques

Current interest detection techniques attempt to probabilistically segment areas of the image pixel-by-pixel [4]. This process could possibly be both simplified and enhanced by

utilizing other image segmenting techniques. For example, Song and Fang [11] have developed a genetic programming (GP) algorithm that detects moving objects in video streams. Segmenting an image by motion would be useful for any mobile robot operating in real-world environments. By segmenting the image in such a way an interest map of positive and negative interests could be generated; since most recognition systems are concerned with inanimate objects, e.g. [2]-[5], areas of movement would have a negative interest while still areas would have a positive interest. Similar schemes could likewise be implemented in dynamic environments where objects of interest could be in motion or at rest.

5.1.4: Alternative Identification Techniques

Each of the five methods in sections 2.3 through 2.7 (pp. 9-13) were selected because of their potential integration with active vision, peripheral-foveal vision or both. For example, by modeling an object by the appearance, spatial relationship and invariants of image patches taken from a sequence of training images, Rothganger et al.'s [6] affine-invariant image patch modeling method seems to lend itself to active vision systems that are driven to seek out novel viewpoints of target objects. At first the limitation that Rothganger et al.'s method does not generalize to the perspective case would seem debilitating for a mobile robot platform, making it unsuitable for a mobile robotic application. However, the authors suggest that this limitation could be remedied by using structure from motion (SFM) or other pose/motion techniques. Each of the above methods could likewise be applied with the use of imagination and a combination of computer vision approaches.

5.2: The Future

Often times, active vision and peripheral-foveal vision go hand-in-hand (e.g. [5]). In the future these methods will likely become further entwined, resulting in highly-integrated and sophisticated vision systems. The hardware (i.e. camera systems) will likely remain varied and (at times) domain-specific, but we can expect to see more novel and unconventional camera systems (e.g. catadioptrics) employed in the future.

The fact that recognition systems strive for the "whole picture"—negotiating around occlusion and selecting appropriate zoom levels to encompass the object of interest in the camera's field of view—we can also likely expect object exploration algorithms in the future to explore their environments in even greater detail. Active camera behavior similar to systems by Farshidi, Sirouspour, and Kirubarajan [3] and Yabuta, Mizumoto, and Arii [9] will become more sophisticated in their ability to select the optimum viewpoint by scrutinizing the scene.

6: Concluding Remarks

Object recognition in real-world environments is a challenging research field. In the case of mobile robotics, the robot's mobility is a valuable tool to be leveraged. Rather than passively examining images to perform perceptual tasks (e.g. object detection, tracking, identification, pose estimation, etc.), the robot's ability to act based on current knowledge can not only simplify the task but improve the quality of information collected for use in that task.

Active vision and peripheral-foveal vision have been successfully used to improve object recognition, though the demonstrated robustness and recognition rate is still far from optimum. The use of scale invariant feature points, particularly SIFT, is commonly used in these systems but research into viewpoint invariant feature points show promise for future implementations. Image segmentation to isolate areas of interest, thus reducing computational complexity and feature mismatching, is commonly employed. New image segmentation (Browning and Veloso's adaptive threshold, and Song and Fang's detector) and feature correspondence techniques (Yabuta, Mizumoto, and Arii's active vision method) could be combined with current research for improved segmentation, reduced computational complexity and/or reduced feature matching errors.

New areas of research have been proposed, namely an omnidirectional camera system for object detection and tracking used in conjunction with active vision and/or peripheral-foveal vision systems; the use of polynocular peripheral-foveal camera systems; and the integration of approaches surveyed in this paper.

Finally, speculations as to the future trends in this area of research were presented. The trend of combining different approaches to gain the advantages of each approach and also compensate for the disadvantages of those approaches will likely continue. Algorithms designed to explore the scene and maximize information for recognition are expected to become more inquisitive as computation speeds increase and camera technology improves.

The contributions of this paper are: an overview of the state-of-the-art in active vision and peripheral-foveal vision research, five state-of-the-art object recognition techniques and their potential contributions to object recognition on a mobile robotic platform, four areas of exploration and future research, and observations on the current and future trends of the research.

References

[1] L.G. Shapiro and G.C. Stockman, *Computer Vision*. New Jersey: Prentice-Hall, 2001, pp. 499-501, 513-517, 548-561.

[2] G. Kootstra, J. Ypma, and B. de Boer. "Active Exploration and Keypoint Clustering for Object Recognition," in Proc. IEEE Int. Conf. on Robotics and Automation, 2008, pp. 1005-1010.

[3] F. Farshidi, S. Sirouspour, and T. Kirubarajan. "Active Multi-camera Object Recognition in Presence of Occlusion," in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 2005, pp. 2718-2723.

[4] S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Meissner, G. Bradski, P. Baumstarck, S. Chung, and A.Y. Ng. "Peripheral-Foveal Vision for Real-time Object Recognition and Tracking in Video," in Proc. Twentieth Int. Joint Conf. on Artificial Intelligence, 2007, pp. 2115-2121.

[5] P.-E. Forssén, D. Meger, K. Lai, S. Helmer, J.J. Little, and D.G. Lowe. "Informed Visual Search: Combining Attention and Object Recognition," in Proc. IEEE Int. Conf. on Robotics and Automation, 2008, pp. 935-942.

[6] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. "3D Object Modeling and Recognition Using Affine-Invariant Patches and Multi-View Spatial Constraints," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2003, pp. 272-277, vol. 2.

[7] W. Shang, X. Ma, and X. Dai. "3D Objects Detection with Bayesian Networks for Vision-Guided Mobile Robot Navigation," in Proc. 8th Int. Conf. on Control, Automation, Robotics and Vision, 2004, pp. 1134-1139, vol. 2.

[8] S. Romdhani and T. Vetter. "3D Probabilistic Feature Point Model for Object Detection and Recognition," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2007, pp. 1-8.

[9] Y. Yabuta, H. Mizumoto, and S. Arii. "Binocular Robot Vision System with Shape Recognition," in Proc. SICE-ICASE Int. Joint Conf., 2006, pp. 5002-5005.

[10] B. Browning and M. Veloso. "Real-Time, Adaptive Color-based Robot Vision," in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 2005, pp. 3871-3876.

[11] A. Song and D. Fang. "Robust Method of Detecting Moving Objects in Videos Evolved by Genetic Programming," in Proc. 10th Annual Conf. on Genetic and Evolutionary Computation, 2008, pp. 1649-1656.

[12] C.G. Fernando, R. Munasinghe, and J. Chittooru. "Catadioptric Vision Systems: Survey," in Proc. Thirty-Seventh Southeastern Symp. on System Theory, 2005, pp. 443-446.

[13] J. Kim, K.-J. Yoon, J.-S. Kim, and I. Kweon. "Visual SLAM by Single-Camera Catadioptric Stereo," in Proc. SICE-ICASE Int. Joint Conf., 2006, pp. 2005-2009.

[14] U.R. Dhond and J.K. Aggarwal. "Binocular Versus Trinocular Stereo," in Proc. IEEE Int. Conf. on Robotics and Automation, 1990, pp. 2045-2050, vol. 3.