Autonomous Learning of a Spatial Model and Object Recognition with a Pan-Tilt-Zoom Camera

Project Final Report

By
Paul Briant
Gary Chern
Jared Starman

CS223B
Dr. Thrun and Dr. Bradski
March 12, 2004

Abstract

The following presents a method for modeling a wide area scene that is observed by a pan-tilt-zoom camera, and for detecting new objects in the scene. The scene is modeled both by SIFT features, as well as color information from a panoramic image of the entire field of view. The panoramic image is stitched together using SIFT features, and new objects are detected by both a feature based detection and color background subtraction. The feature based detection is based on the appearance and occlusion of SIFT features, while the color based detection is done by comparing histograms of hue values. The use of these two methods in combination is a unique solution that is in general more robust than existing methods. The SIFT features have shown to provide a good basis for stitching the panoramic background image, and the object detection usually produces satisfactory results.

Autonomous Learning of a Spatial Map and Object Recognition with a Pan-Tilt-Zoom Camera

1.0 Introduction

Current surveillance systems require human monitoring, which is time consuming, expensive, and prone to error. Efficient and accurate computer object recognition and tracking algorithms would help overcome this problem and would greatly enhance the security industry. Before object recognition can occur, however, objects must first be detected as foreground. Object detection requires a system to learn a model of its background, so that foreground objects can be detected through simple comparisons between new images taken by the camera and the background model. For static surveillance cameras, learning the background model and object detection are usually straight-forward, however, surveillance systems often employ pan-tilt-zoom (PTZ) cameras, which make these tasks more difficult. The underlying new problem for PTZ cameras is that the camera can only see a portion of its total field of view at any instant in time. Therefore, in order to build a usable background model, the system must stitch together different pieces of its background accurately. In addition, when a new image is taken for detection, the system must correctly position the image before it can be compared to the background model.

Although there has been much work in both panoramic image construction, and in object detection, there are currently few algorithms that combine these two areas for object detection over a wide field of view. In addition, the few algorithms that do exist are based primarily on pixel color information [5, 6]. Our approach differs from these previous algorithms because we use not only the pixel color information, but also Scale Invariant Feature Transform (SIFT) features. Using these SIFT features as well as the color information provides a more robust solution to this problem.

Our approach uses the SIFT features both for building the background model and during object detection. While building the background model, they are used to develop an initial set of corresponding points between a new image and the current background [1, 8]. The initial set of corresponding points is obtained by performing a nearest neighbor search in feature vector space. These corresponding points can then be used to calculate an affine transform between the image and the current background. SIFT features provide a good basis for the affine transform because they are well localized and fairly unique. Once this affine transform is known, the image and its features can be merged with the current background.

After an initial background model is established, object detection can begin. The SIFT features are again used to determine an affine transform between a new image and the background, in order to correctly position the new image in the background. After the image is correctly positioned, the algorithm does both a feature based and a color based object detection. Only those regions that appear as foreground in both detection algorithms are considered candidate foreground. Masking the color detection in this manner helps to reduce the noise. After the foreground is detected, the new image is

incorporated into the background in order to update the background. This allows for a background model that adjusts automatically to long term changes in the scene.

The algorithm for building a background model and stitching together images shows good results. The camera is currently in the Robotics Laboratory in the Gates building, and it can build up a good panoramic image of the laboratory. The stitching algorithm typically has better results in areas with high texture, although, results were still adequate in areas with limited texture. The object detection algorithm usually produces satisfactory results; however this is not always the case.

2.0 Related Work

Object detection over a wide field of view draws upon two fields: creating panoramic images and object detection.

2.1 Creating Panoramas

There are two basic methods of constructing panoramas: direct methods and feature based methods. Direct methods attempt to iteratively minimize an error function over the area of overlap of two images [1]. Feature based methods involve finding corresponding points between images, calculating an affine transform between the points, and using the affine transform to project points from one image to the other. There are different methods of finding corresponding points; they usually involve finding features that match between images. Our method was very similar to the method by Brown and Lowe [1], who uses SIFT features to establish the corresponding points. SIFT features are well suited for this task because they are characterized by 128 long feature vector, which makes each feature fairly unique. This uniqueness allows corresponding features to be determined through a nearest neighbor search over feature vector space.

2.2 Object Detection

The method most widely used for detecting moving objects is background subtraction, which works by subtracting the intensities or color of a background image from the current video frame. If the camera is fixed and always focused on the same scene, the simplest background model is an image of the scene with no new objects in it (stationary background). However, due to changes in illumination and changes in the background scene (i.e. transient background, such as swaying tree branches), the background image is often a model that needs to be updated as frames are gathered from the video stream. Problems facing maintaining the background model include moved background objects, gradual or sudden changes in illumination, vacillating backgrounds like swaying trees, camouflaged new objects that are the same color as the background, and shadows. More potential problems are identified in [2]. Other background subtraction algorithms use statistical information from previous frames to model the background. The color or intensity at each pixel in the image can be modeled by a single Gaussian or a mixture of Gaussians. These approaches allow for variation in the scene due to noise. A table of different background subtraction approaches is summarized in [3].