

# UNDERGRADUATE REPORT

Using Computer Vision to Train a Sound Tracking System

*by Charlie Brubaker, Stephanie Wojtkowski*  
*Advisor: P.S. Krishnaprasad*

**U.G. 2000-2**



*ISR develops, applies and teaches advanced methodologies of design and analysis to solve complex, hierarchical, heterogeneous and dynamic problems of engineering technology and systems for industry and government.*

*ISR is a permanent institute of the University of Maryland, within the Glenn L. Martin Institute of Technology/A. James Clark School of Engineering. It is a National Science Foundation Engineering Research Center.*

**Web site <http://www.isr.umd.edu>**

# Using Computer Vision to Train a Sound Tracking System

Charlie Brubaker and Stephanie Wojtkowski  
Institute for Systems Research

August 4, 2000

## Abstract

In this research, computer vision was used to locate a sound source for feedback into an audio system. The camera was first calibrated to determine the relationship between the world coordinates and the pixel coordinates of an object. To aid in the calibration process, computer vision techniques such as gradient calculation and the Hough Transform were used to extract the calibration points from a series of images. These points, along with their corresponding world coordinates, were then used in Roger Tsai's camera model to calibrate the camera.

The intrinsic and extrinsic camera parameters were then used to find the vector of the sound source in an image. Again, vision processing was used to extract the sound source from an image using red as a detectable feature. The largest red region was isolated, and the centroid of that region was used to mark the location of the sound source. Finally, Tsai's model was used in reverse to find the vector in the world along which the camera lies.

## 1 Introduction

This project was intended to complement a sound tracking system produced by the Neural Systems Laboratory at the Institute for Systems Research. The purpose of this sound tracking system is to determine the location of a sound source given the inputs from two microphones. The answer to this problem is not straightforward, particularly if there is an obstruction in front of one or both of the microphones. In this case, it would be quite difficult to produce an equation for the location of the sound source.

Instead of trying to solve this problem explicitly, the researchers at the Neural Systems Laboratory have chosen to use a learning algorithm to train the sound system. As with all learning algorithms, however, it is necessary to evaluate the accuracy of the system's output after each run. Instead of measuring each location of the sound source explicitly, a vision system was created to determine the location of the sound source autonomously.

To construct a vision system for this purpose, the camera was first calibrated using Roger Tsai's camera model. An image is taken by the system, and the sound source is located in the image. The calibration information is then used to determine the vector along which the sound source lies.

## 2 System Specifications

The sound system consists of a Styrofoam head with two microphones mounted at the approximate location of human ears. Pinnae will be added to create a more realistic model.

The robot used for this project was the Super Scout II, a product of Nomadic Technologies. This model includes bump sensors, 16 sonar sensors, odometric sensors, and a Pentium 233 MHz processor.

The camera selected for the vision system was an Xcam 2, manufactured by X10. This camera was selected for a couple of reasons. First, this camera allows us to demonstrate that sophisticated vision sensing can be performed with a low quality sensor. Second, this model is wireless, so all vision processing can be executed off-board.

The framegrabber used in this system was a Matrox Meteor II. The framegrabber was installed on the Windows NT machine on which processing was performed.

### 3 Calibration

Before an image can be used to determine the location of an object in the world, the camera must be calibrated. This calibration determines the intrinsic and extrinsic camera parameters so that an accurate transformation can be performed between image coordinates and world coordinates. The calibration method chosen for this system was Roger Tsai's camera model.

#### 3.1 Roger Tsai's Method

The Tsai camera model [1] uses six extrinsic and five intrinsic parameters to describe the transformation from world coordinates to a pixel in the frame buffer. The six extrinsic parameters, yaw, pitch, roll,  $T_x$ ,  $T_y$ , and  $T_z$ , describe the transformation from the world frame to the camera frame, in which the  $x$  and  $y$  axes form the image plane. The focal length then determines the perspective projection from the camera coordinates to image coordinates. The radial lens distortion can be described by an infinite series, but Tsai's model only keeps the linear term, whose coefficient is  $\kappa_1$ . Applying this transformation gives distorted image coordinates. To account for displacement of the CCD and errors in the signal processing of the frame grabber, Tsai introduces the parameters  $C_x$ ,  $C_y$ , and  $s_x$ , the uncertainty scale factor in  $x$ .

The calibration of these eleven parameters consists of collecting a data set containing world coordinates and the position of the corresponding pixel in the frame buffer. The key step in Tsai's calibration method is setting up and solving an overdetermined system of linear equations. For details see his paper.

#### 3.2 Automated Point Detection and Correlation

By far, the most tedious parts of camera calibration are the measurement of each calibration point in the world coordinate frame and the extraction of each point from an image, both of which are done by hand. While there is no way to avoid measuring the world coordinates of each point by hand, vision processing routines can be written to locate the calibration points automatically and then properly correlate them with the appropriate set of world coordinates. The process that was used for this system is described below.

The module takes in a file containing the names of files in which the world coordinates are written, and the image that corresponds to each file. The calibration is performed using a flat board on which a broad grid is drawn. The intersections of the grid lines will be used as the calibration points, since they are easy to locate. Images are taken at several distances from the board to produce a 3-dimensional cloud of calibration points. The calibration images would look like those in Figure 1.

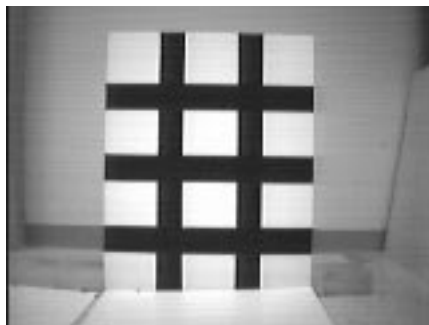


Figure 1: Calibration Image

### 3.2.1 Grid Extraction

Each image is then processed using computer vision techniques to extract the calibration points. The first step in this processing is the application of a gradient filter. The filter calculates the intensity difference between the current pixel and each adjacent pixel. If this difference is above the predetermined threshold, the pixel is considered to contain a high gradient and is colored white to distinguish it from the background. An array containing the magnitude of the highest gradient at each pixel location and the direction of that gradient is retained. This array is used to thin the gradient image in the direction of the gradient using a Canny operator. For each white pixel, the gradients of the adjacent pixels are referenced in the gradient array. If one of the adjacent pixels has a higher gradient, the current pixel is changed to a background pixel. Otherwise, the adjacent white pixels are changed to background pixels. Figure 2 shows an example of the resulting image.

### 3.2.2 Line Extraction

A method called a Hough Transform is performed on each white pixel in the gradient image. This technique extracts the lines from the gradient image. To perform a Hough transform, the space of all lines that could intersect with the image is discretized based on the angle and radius of those lines from the center of the image (the origin). Then, each white pixel "votes" for all of the lines that could cross it. The angle and radius combinations with the most votes

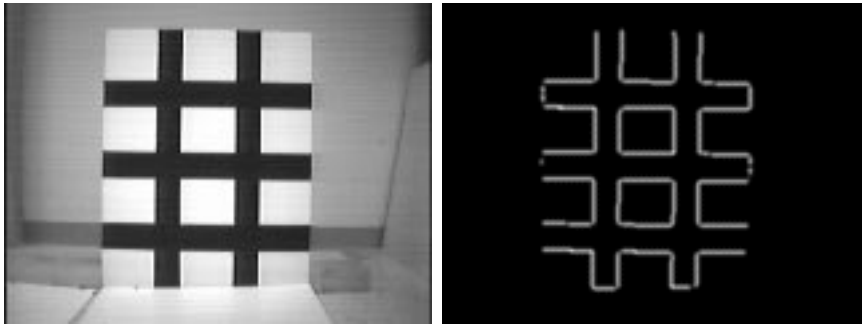


Figure 2: Edge Detection

should correspond to the lines in the image. In a graphical representation of the hough space, it is evident that there are ten strong maxima in the space, each one corresponding to one of the lines in the image above. To extract only those lines in the image that received the most votes, a non-maxima suppression was performed on the Hough space.

Given these ten lines, their intersections can be found. These intersections are the calibration points. Figure 3 is a composite image containing a copy of the original image and a copy of the gradient image with the locations of the lines (green) and the location of each calibration point (red) drawn in.

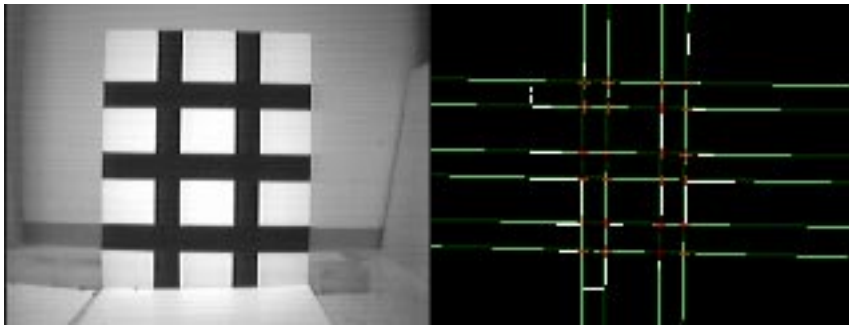


Figure 3: Finding the Lines and Intersections

It is important to note that the grid lines need to be of significant thickness, or the images need to be high resolution. The images in Figure 4 were taken using a grid with 1 inch lines. Because of inadequate resolution in the calibration images, the black/white gradient was not as high as was necessary, and imprecise identification of the calibration points resulted.



Figure 4: Bad Calibration Image

## 4 The Object Recognition System

The object recognition system receives the sound system's estimate of the angle from which a sound came. First, an image is taken of the current scene. The image is then subjected to vision processing to extract the sound source. Tsai's method is then used in reverse to determine the vector along which the centroid of the sound source lies in the world. The azimuth and elevation of the sound source are returned to the sound module.

### 4.1 Locating the Sound Source

#### 4.1.1 Thresholding

The feature that was chosen to distinguish the sound source from the background was the color red. A pixel was selected as a part of a red region based on two criteria. First, the ratio of the red intensity of the pixel to the overall intensity of the pixel had to be above a certain threshold. The overall intensity was defined as the sum of the red, green, and blue intensities, each of which ranged from 0 to 255. This criterion ensured that red was the predominant color of the pixel. Second, the red intensity of the pixel had to be higher than a certain threshold. This criterion was added to make sure that black pixels, which have very low intensity values, were not included in the selected pixels, even if the black pixels were predominantly red. Figure 5 shows the original image and the thresholded version of this image. Note that the darker shades of red in the object were not extracted. Both thresholds must be adjusted to accurately extract the target object.

#### 4.1.2 Growing and Shrinking

The thresholded image can be very rough due to noise and small red objects in the image. Growing and Shrinking is a technique that removes such noise from the image and makes the object region more contiguous. First, each region is grown several times. Growing involves adding each background pixels that is



Figure 5: Thresholding to Find Red

touching an object pixel to the object. The object literally grows larger as holes are filled in. Next, the object is shrunk an equivalent number of times to reduce it to its normal size. The finished product is an image containing object regions with smoother boundaries and less noise. Figure 6 shows the before and after images of this process.

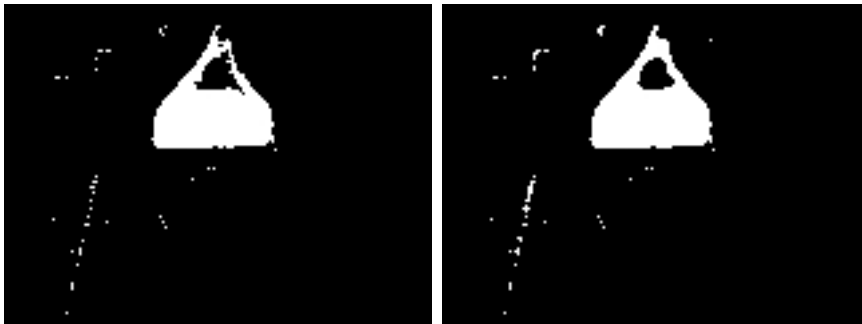


Figure 6: Grow and Shrink to Consolidate Regions

#### 4.1.3 Connected Region Extraction

Once the sound source has been extracted from the image, the x and y positions that contain the most white pixels could be used as the center of the object. This technique is fine if sound source is the only red object in the image, but if there is more than one red object, or if the image is noisy, this calculation can be significantly displaced from the actual center of the object.

To solve this problem, a connected region extraction was first performed on the thresholded image. A connected region extraction gives each region in the image its own unique identification number, so that one region can be distinguished from another. The algorithm first looks for an object pixel that is not yet part of a region. Once it finds one, the pixel is marked as part of a



new region, and all connected object pixels are recursive marked as members of the same region. Figure 7 shows an image with different gray levels for each region.

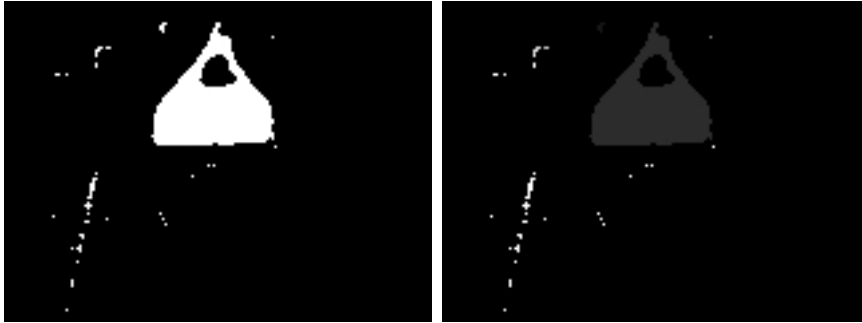


Figure 7: Connected Region Extraction

#### 4.1.4 Largest Region and Centroid Location

While the connected region extraction was being performed, the program kept track of the number of pixels that was added to each region in the image. The identification number of the region with the maximum number of pixels could then be found easily by searching this array. The largest red region in the image was assumed to be the sound source; all others were considered to be smaller red objects or noise.

Once the identification number of the largest region was located, the image was traversed one last time. This time, the x and y values of the pixels that were members of the largest region were summed and used to calculate the centroid of the largest region. This centroid could then be used to calculate the vector of the sound source in world coordinates. Figure 8 shows an image with the centroid of the largest region indicated.



Figure 8: Finding the Centroid

#### 4.1.5 Conversion to World Coordinates

Now that the location of the sound source in the image had been calculated, Tsai's model could be used to find the real world vector along which the sound source lay. To convert from the position of a pixel in the frame buffer to world coordinates, we need only retrace the steps of Tsai's camera mode. All steps are invertible with the exception of the projection. Here we assume an arbitrary positive z value (distance from the image plane) to obtain a vector in 3-D camera coordinates. The length of the resulting vector may not be correct, but the direction is. We then rotate the vector back to the world frame. The result is a vector from camera pointing toward the object identified in the image.

## 5 Future Work

There are two main areas in which this project could be expanded. First, sound and vision could be used together to get a better estimate of the position of the sound source. Second, a method such as elliptical head tracking could be used to get the system to follow a speaking person instead of a red object.

## References

- [1] Roger Y. Tsai. "A versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses", *IEEE Journal of Robotics and Automation* , Vol. RA-3, No. 4, August 1987, pages 323-344.