Visual Target Segmentation and Identification

Lana Haru Carnel
Electrical Engineering
University of Tennessee
lcarnel@utk.edu

Graduate Mentor: **Parvez Ahammad**
Research Supervisor: **Dr. Jonathan Sprinkle**
Faculty Mentor: **Prof. S. Shankar Sastry**

July 29, 2005

Summer Undergraduate Program in
Engineering at Berkeley (SUPERB) 2005

Department of Electrical Engineering and Computer Sciences
College of Engineering
University of California, Berkeley

Visual Target Segmentation and Identification

Lana Haru Carnel

Abstract

Locating, isolating and tracking the objects of interest are a key step in visual scene analysis in surveillance. Reflection, variations in appearance, clutter, and occlusion create challenges in identifying the object of interest robustly. Cues (or features) such as color, motion, size and dynamics must be used in tandem, to effectively decide what is relevant and discard the unnecessary information based on the goal at hand. In this work, we look at the target identification and tracking problem from the point of view of building surveillance applications and address the aforementioned issues by using multiple layers of segmentation. The approach involves intelligent feature selection and robust combination of these features to locate and track these objects. We demonstrate the results of our implementation on videos and images taken both in controlled and uncontrolled environments.

1 INTRODUCTION

Surveillance systems are systematically used to detect an anticipated set of targets. There are existing networks of cameras currently in place in commercial businesses, health care institutions, offices, and schools. The monitoring of these networks is currently conducted by humans. The implementation of computer vision will make these existing networks more efficient and economically feasible. Camera vision encompases the idea that cameras can begin to mimic the process of human sight to analyse the visual information the network receives. It is through this concept that a camera ceases to be merely a method of documentation and becomes a tool that can process intent. Using existing networks in uncontrolled environments requires the consideration of factors including: challenges with existing lighting, occlusion, and reflection. Ultimately, computer vision for this application will take the feed from each camera, perform analysis of each image, outputting the coordinate movements for each target and place them in a global framework with respect to the monitored geography.

Implementation of analysis in camera vision begins with separating the object of interest from the unecessary information. Once this separation has been made, the targets must be distinguished from one another. Data must be then be outputted in a format for further analysis. These steps are important as they act as the liason between the acquisition of image and the development of environmental awareness. Our work is concerned only with constructing an algorithm for these intermediate steps with respect to surveillance types of scenarios in both static(controlled) and moving(uncontrolled) environments (Fig. 5). Output for this work exists in two forms: placing a marker on the image within the video and listing the local coordinate locations of the center of the target within the video frame.

1.1 Past Work

Applications for computer vision are currently moving toward the recognition of human movement. These types of systems are used in a wide variety of platforms. The use of cameras to analyse a wide scene requires the ability to articulate separate movements when the indicators have relatively small magnitudes [1]. Detection algorithms in surveillance applications where resolution is very low requires the implementation of combinations of features such as motion and intensity [5].

The idea of pulling relevant information out of a visual scene is most closely informed by the method of human sight. The eye and brain use hundreds of visual features to discern pertinent information in a scene. To mimic this highly robust system, we approach a video image with a systematic way of checking features. The segmentation algorithm itself is of primary importance in this process. This concept is implemented on a much smaller scale in this project to effectively reduce the items of interest to a single coordinate in order to track its location. High level behavior like human motions has been examined by segmenting the body and matching movement with a predetermined hierarchy to determine movement [4]. Additionally, data driven approaches learn categories of action events from the given data to classify motion [6]. These high level analysis systems use an initial segmentation process of masking desired information from the initial information. This segmentation has historically required specialized segmentation processes for the given situation. It is this aspect of image processing that we examine in this work.

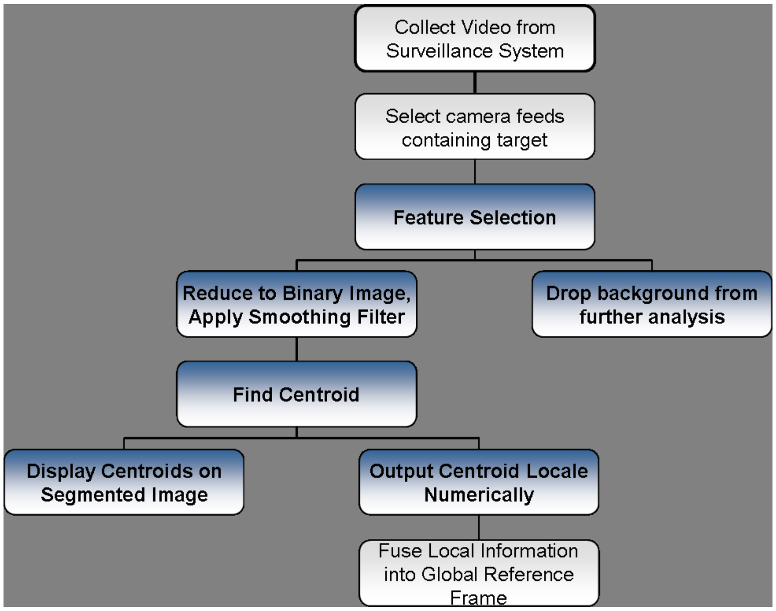
Collect Video from Surveillance System
Select camera feeds containing target
Feature Selection
Reduce to Binary Image, Apply Smoothing Filter
Drop background from further analysis
Find Centroid
Display Centroids on Segmented Image
Output Centroid Locale Numerically
Fuse Local Information into Global Reference Frame

Figure 1. Proposed data flow algorithm

2 IMAGE ACQUISITION

2.1 Existing Camera Network

Image acquisition was based around an existing camera network monitoring the third floor hallway of Cory Hall at the University of California, Berkeley. The system is constructed with two omnidirectional (Figure:2)and fourteen unidirectional cameras(Figure:3). Omnidirectional cameras are placed in the two corners of the monitored areas with the unidirectional cameras. The omnidirectional cameras appear as the two circular images in (Figure:4). Together they encompass the entire monitoring area, while the unidirectional cameras provide a more detailed and confined image.

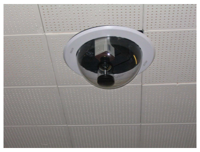

Figure 2. Omnidirectional Camera

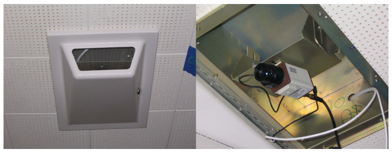

Figure 3. Unidirectional Camera

2.2 Still Images

Data sets for this work were manufactured to provide the desired conditions for testing. Both still and video images were collected to develop and test the analysis methods. The environment was chosen for its existing camera network, accessibility, and representation of visual challenges. The targets were established by choosing a color, bright pink, that was unique to the testing environment.

Still photographs were staged by placing the target(s) in the desired environment with uniform lighting conditions, minimal reflection and no occlusion. Two images best representing this controlled environment were chosen. A single target is shown in one image; a pair of targets are shown in the second.

2.3 Moving Image

Images were collected from the existing camera network. Motion was introduced by attaching the targets to remote controlled cars. A video clip was chosen that provided several challenges to the segmentation process. Because of the existing lighting, the targets themselves vary greatly in color as they move within the camera frame. The floor and walls are highly reflective providing additional challenge. The clip represents the presence of zero, one, two, or three targets in each frame. The distinction of the targets are further challenged by occlusion created by people moving in and out of frame, and the merging of the two targets as they 'bump' into each other in the hallway.Testing was first completed on the still images and then adapted to meet the challenges provided in the video clip.

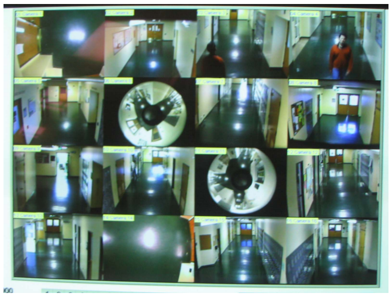

Figure 4. Video Display of All Cameras

3 IMAGE SEGMENTATION

3.1 Feature Selection

Of primary concern in this project is segmentation by color. The target maintains a unique color in both environments, providing a base to begin segmentation. The representation of color in digital formats. Two standard color models were examined for use. RGB (Red, Green, Blue) and HSV (Hue, Saturation, Value). Colors in both spaces are determined by the combined values of three layers. RGB color is constructed using three monochromatic color layers and can be described in Cartesian Coordinates(Figure:5). HSV is constructed using sight based components of color, decoupling the intensity component from the color-carrying information [2]. Cylindrical coordinates are best used to describe these values(Figure:6).

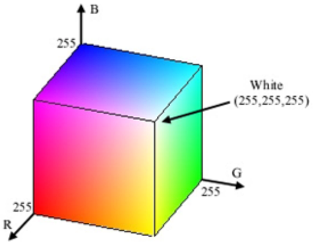

Figure 5. Geometric representation of RGB color

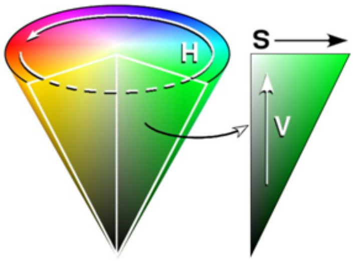

Figure 6. Geometric representation of HSV color

Both RGB and HSV color space based segmentation were impemented using the still images. Effective segmentation of the RGB image required using parameters for each color layer combined with a weighted tolerance for each. The HSV image was effectively segmented using only one color layer and tolerance(Figure:3.1). Despite our expectation that the hue value would be most distinctive within the frame, the saturation layer produced the most effective.

Using the HSV color space, the moving images were then segmented. Due to the increased complexity within the image, it was necessary to use the additional two layers to separate the target. Even with the additional parameters, noise was still evident within the frame.

The resulting images were constructed and stored as a binary image. Pixels meeting the three parameters within the tolerance receive the value '1' and any pixels not meeting these requirements were labeled '0'. During this process, the image is flattened from three layers to one.

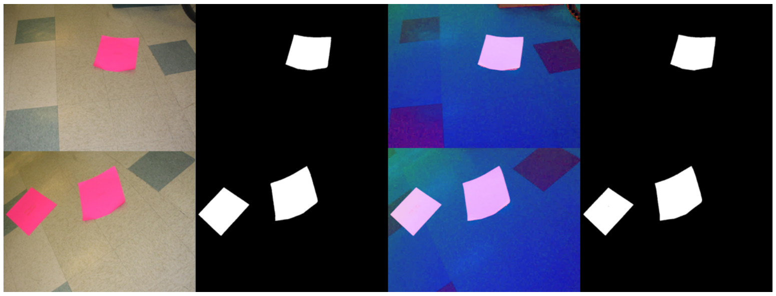

Figure 7. RGB segmentation using 3 parameters; HSV segmentation using 1 parameter

3.2 Smoothing Filter

The presence of noise around target edges and stray pixels after segmentation require additional measures to prepare the image for further analysis. We remove this noise by applying a smoothing filter to remaining binary image. The smoothing filter is implemented by applying a moving point average within a 5×5 matrix where $h\left[x, y\right]$ is the new pixel value computed by the average of the surrounding pixel values of $g\left(x, y\right)$.

$$h\left[x, y\right] = \frac{g\left(x-2, y-2\right) + g\left(x-2, y-1\right) + g\left(x-2, y\right) \ldots + g\left(x+2, y+2\right)}{25}$$

The resulting image can now be thought of as a grayscale image varying in value from 0.0 to 1.0. This minimizes the impact of stray pixels. This also has a negative impact on the boundaries of the target. As these edges maintained sharp contrast in the binary image, the smoothed image blurs these edges, changing the perimeter shape slightly.

Once new pixel values are computed, they are reconverted to a binary image by subjecting the new values to a tolerance range:

$$h\left(\left[x, y\right] \leq .1\right) = 0$$

$$h\left(\left[x, y\right] > .1\right) = 1$$

The tolerance distribution heavily favors returning the '1' value in the interest of removing stray pixels, rather than reducing the size of the targets.

4 IDENTIFICATION

Once the object(s) of interest have been effectively separated from other elements, it is then necessary to represent the target in a usable format. Identification of the targets requires two components: distinguishing multiple targets within the same frame and establishing the objects location.

4.1 Clustering for Multiple Targets

The appearance of a single target within the frame is easily dealt with. Using the binary format, all pixels valued at '1' can be considered as indicators of the single object and no further steps are necessary to distinguish it. When multiple targets are present, further considerations must be made. Groups of pixels are then distinguished by checking pixels with value '1' for connectivity. This is implemented by checking each pixel with its eight adjacent pixels. The edges of the group are defined when none of the adjoining pixels are equally valued. Each new group is then revalued to make it distinct. When considering each target individually, its value is returned to '1' to preserve the conditions of the binary image.

4.2 Centroid Calculation

Once the targets have been uniquely defined, they must be presented in an effective means for further analysis. Because the objects' size changes dramatically as it moves toward the horizon point, an accurate representation of its location is by its central point. By defining this point on each target in each frame, location can be outputted numerically by its 'x' and 'y' coordinate value with respect to the image plane. This idea can be mathematically described by examining the center of mass of a lamina[3] with $\sigma\ (x, y)$ equal to one, as the surface density of each target is constant:

$$M = \int\int \sigma(x, y)dA$$

$$\bar{x} = \frac{\int\int x\sigma(x, y)dA}{M}$$

$$\bar{y} = \frac{\int\int y\sigma(x, y)dA}{M}$$

$$\text{centroid} = (\bar{x}, \bar{y})$$

The center of mass or centroid most accurately measures its relational location between frames because it is a representation of the mean of each dimension, and it does not fluctuate as readily as the boundaries of each object. This calculation is visually represented on top of the binary image as a blue asterisk indicating the centroid of each target.

5 RESULTS

The algorithm successfully finds the targets and outputs the centroid. These results are shown using a selection of frames from the video clip. The original, unprocessed image is shown to the left, with its corresponding processed counterpart to the right. The results of feature selection and smoothing appear as white; the centroid is denoted with a blue asterisk(Figure:5).

5.1 Limitations

The nature of our feature selection is limited to the targets with distinctive color within the surveillance environment. The segmentation and identification of targets closely matching the background may cause difficulty. It would be necessary to then use a weighted heirarchy of feature selection to effectively separate the image. This algorithm does not distinguish between targets when they appear to merge within the video frame. This problem could be corrected by maintaining the boundary of the shapes for identification, rather than the cententroid. Finally, the location output examines the visual information with respect to a two dimentional plane. If the elevation of the target changes dramatically within the environment, the coordinate output will not be a direct reflection the object's location in a global sense. This problem can be solved using multiple cameras to create three dimentional identification. Alternatively, relative size of the target can be established with respect to its position on the floor. Assesment can then be made of its true location by analysis of the target's size and location within the image frame.

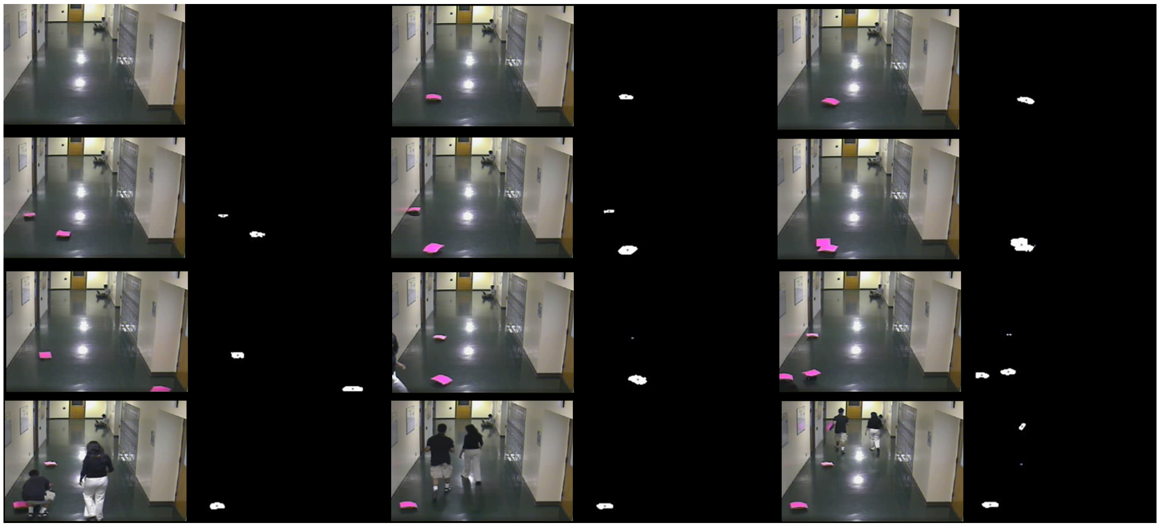

Figure 8. Selected Video Frames and Resulting Segmentation and Identification

5.2 Integration with Camera Vision

Within our proposed surveillance application, this work can be augmented by applying it to each camera in the network. After the segmentation and identification alorithm is complete, output data from the centroid location would be calibrated to transform the frame location coordinates to the larger coordinate system of the floor plan. The trajectory of the targets' movement can now be computed with respect to actual location, and the implications of these movements can be made. The unprocessed output and the segmented image can then be shown alongside this global map to provide additional detailed visual information.

5.3 Future

Real world application of this system may lie in augmenting many types of pre-existing surveillance systems. With camera networks already in place in commercial and public arenas, this technology can be applied to the specific requirements of each system. The addition of camera vision to these networks can eliminate the neccessity of human evaluation of video, ultimately creating a reduction in time and cost of surveillance analysis.

6 Conclusion

We have developed an algorithm for the segmentation and identification of targets from visual data. Using color as a primary feature selector, the system can effectively determine the target(s) location within the image frame. Results using a video created with a surveillance camera demonstrate its effectiveness in handling conditions typical of shared school and commercial interiors.

CKNOWLEDGMENTS

This work would not have been possible without the help of Andrew Chekerylla. Thank you John Suarez for being the sounding board and company for the late night pizza runs.

References

[1] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[2] R. Gonzalez and R. Woods. *Digital Image Processing*. Prentice Hall, New Jersey, 2002.

[3] B. K. P. Horn. *Robot Vision*. MIT Press, Cambridge, MA, 1986.

[4] N. T. Nguyen, H. H. Bui, S. Venkatesh, and G. West. Recognising and monitoring high-level behaviours in complex spatial environments. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[5] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *International Conference on Computer Vision*, 2003.

[6] H. Zong and J. Shi. Finding (un)usual events in video. Technical report, Robotics Institute, Carnegie Mellon University, May 2003.