UNIVERSITY OF MINNESOTA

CENTER FOR TRANSPORTATION STUDIES

INTELLIGENT TRANSPORTATION SYSTEMS INSTITUTE

Recognition of Human Activity in Metro Transit Spaces

Final Report

Prepared by
Guillaume Gasser
Nathaniel Bird
Nikolaos Papanikolopoulos

Department of Computer Science and Engineering
University of Minnesota

CTS 04-02

HUMAN CENTERED TECHNOLOGY TO ENHANCE SAFETY AND TECHNOLOGY

Technical Report Documentation Page

1. Report No.	2.	3. Recipients Accession No.
CTS 04-02		

4. Title and Subtitle	5. Report Date
RECOGNITION OF HUMAN ACTIVITY IN METRO TRANSIT SPACES	June 2004
	6.

7. Author(s)	8. Performing Organization Report No.
Guillaume Gasser, Nathaniel Bird, Nikolaos Papanikolopoulos	

9. Performing Organization Name and Address	10. Project/Task/Work Unit No.
University of Minnesota Department of Computer Science and Engineering 200 Union Street SE Minneapolis, MN 55455	11. Contract (C) or Grant (G) No.

12. Sponsoring Organization Name and Address	13. Type of Report and Period Covered
Minnesota Department of Transportation 395 John Ireland Boulevard Mail Stop 330 St. Paul, Minnesota 55155	Final Report
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract (Limit: 200 words)

In this report, we introduce a vision-based system to monitor for suspicious human activities at a bus stop. The system currently examines behavior for drug dealing activities which is characterized by individuals loitering around the bus stop for a very long time with no intention of using the bus. To accomplish this goal, the system must measure how long individuals loiter around the bus stop. To facilitate this, the system must track individuals from the video feed, identify them, and keep a record of how long they spend at the bus stop. The system is broken into three distinct portions: background subtraction, object tracking, and human recognition. The background subtraction and object tracking modules use off-the-shelf algorithms and are shown to work well following people as they walk around a bus stop. The human recognition module segments the image of an individual into three portions corresponding to the head, torso, and legs. Using the median color of each of these regions, two people can be quickly compared to see if they are the same person.

17. Document Analysis/Descriptors		18.Availability Statement
Monitoring Human Activity Transit Space	Event Detection Camera Loitering Activity	No restrictions. Document available from: National Technical Information Services, Springfield, Virginia 22161

19. Security Class (this report)	20. Security Class (this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	18	

RECOGNITION OF HUMAN ACTIVITY IN METRO TRANSIT SPACES

Final Report

June 2004

Principal Investigator:
Guillaume Gasser
Nathaniel D. Bird
Nikolaos P. Papanikolopoulos

Artificial Intelligence, Robotics and Vision Laboratory
Department of Computer Science and Engineering
University of Minnesota

Center for Transportation Studies
University of Minnesota

CTS 04-02

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

In this report, we introduce a vision-based system to monitor for suspicious human activities at a bus stop. The system currently examines behavior for drug dealing activities which is characterized by individuals loitering around the bus stop for a very long time with no intention of using the bus. To accomplish this goal, the system must measure how long individuals loiter around the bus stop. To facilitate this, the system must track individuals from the video feed, identify them, and keep a record of how long they spend at the bus stop. The system is broken into three distinct portions: background subtraction, object tracking, and human recognition. The background subtraction and object tracking modules use off-the-shelf algorithms and are shown to work well following people as they walk around a bus stop. The human recognition module segments the image of an individual into three portions corresponding to the head, torso, and legs. Using the median color of each of these regions, two people can be quickly compared to see if they are the same person.

We show that the system can successfully track individuals in sparsely-populated outdoor scenes with limited occlusion in near real time. We also show that individuals can be correctly distinguished from each other with an accuracy of about 80 percent.

CHAPTER 1

INTRODUCTION

Overview

The purpose of this work is to develop a vision-based system that monitors the activities of individuals at a bus stop for suspicious behavior. Autonomous vision-based systems are ideal for monitoring human activities in public places such as bus stops because they are more "attentive" than a human and they free up manpower that is better assigned elsewhere.

Focus is placed on monitoring for behavior indicative of drug dealing. According to officials at Minnesota's Metro Transit, the central behavior associated with drug dealing is presence at a bus stop for extended periods of time, indicating the person in question is loitering as opposed to taking the bus. It is important to note that drug dealers loitering around a bus stop can leave periodically and come back later, making it important to keep a record of people who have spent a lot of time at the bus stop recently and check if they have come back. Because of this, it is not sufficient to use only motion tracking to keep track of how long a person has been in the scene in order to accurately time how long they have been loitering around the bus stop. An additional procedure to recognize that a given person has been seen before must be implemented as well.

Previous work

To achieve our goal, our system employs techniques for foreground segmentation, tracking, and recognition. In this section, we discuss some of the relevant research pertaining to these areas.

Research has already been done in the field of segmentation. Prior methods for motion segmentation such as static background subtraction work fairly well in constrained environments. But these methods are not suitable for unconstrained, continuously changing environments like outdoor scenes. So it is important to find a statistical way to model the color of each pixel that can work even with unconstrained scenes. One of the simplest methods, introduced by Wren *et al*. [1], is to model the intensity of each pixel by a single Gaussian. This works well in relatively static indoor

environments. Alternatively, Friedman and Russel [5] used a mixture of three Gaussians for each pixel using an incremental maximization method. Stauffer and Grimson [2] used a mixture of Gaussians for each pixel to adaptively learn the model of the background. Nonparametric kernel density estimation has been used by Elgammal *et al.* [3] for scene segmentation in complex outdoors scenes. The later method gave the best results for the purposes of the bus stop monitoring system.

There has also been a plethora of research into the area of vision-based tracking. For example, multi-level tracking has been used in Cucchiara *et al.* [6] for monitoring traffic. McKenna *et al.* [7] performed three-level tracking consisting of regions, people, and groups in indoor and outdoor environments. Kalman filter-based feature tracking for predicting trajectories of humans was implemented by Rosales and Sclaroff [8]. Koller *et al.* [9] used a tracker based on two linear Kalman filters, one for estimating the position and the other for estimating the shape of the vehicles in a highway scene. Some other tracking methods are based on the color distribution of the target and not on position prediction through a Kalman filter. This is the case for the method developed by Comaniciu *et al.* [4], in which the new target position is found by searching in the target's neighborhood in the current frame and computing a correlation score, the Bhattacharyya coefficient.

The problem of identifying humans from video in controlled environments is quite challenging. The problem becomes further exacerbated when the video is of an outdoor scene and when humans are distant from the camera, occupying a small area within the image. Not much research has dealt with all these complexities in the past. Previous research into visual recognition deals with recognizing objects and actions in very constrained, structured environments. Nayar *et al.* [10] introduce a system that first creates a library of images for each object to be recognized by taking pictures of it from many different angles. The model formed from this library of images is then shown to be able to recognize the object from any novel angle. This is performed in a controlled, indoor environment on rigid objects. Elgammal *et al.* [3] utilized a color-density-based image segmentation method to aid in the location of people within a video segment by locating color "blobs" relating to the head, torso, and legs of a person.

To identify specific actions, Efros *et al.* [11] introduce a system that compares the optical flow pattern in a novel video of a person performing an unknown action to a database of optical flow patterns for known actions. A matching algorithm is used to determine whether both videos show people performing the same action. This is shown to work decently in specific outdoor environments devoid of shadows and significant forms of occlusion. This method is also limited by the scope of its action database but seems promising for identifying well-defined behaviors.

Implementation

There are many difficulties to overcome when implementing a vision system to work in unconstrained environments such as the outdoors. A typical frame from a video of a bus stop can be seen in Figure 1. As this scene illustrates, the system is intended for outdoor use. Therefore, a wide range of possible lighting conditions must be accounted for. Direct sunlight, cloudy conditions, and nighttime are among the possible illumination types that will be present in an outdoor environment. Another obstacle to overcome is the existence of shadows, caused either by the sun or by artificial light sources at night. Occlusion must also be accounted for. Unmovable obstacles such as street signs, newspaper machines, fire hydrants, and the bus stop itself can all block the view of a given individual in the scene. Also of concern are occlusions of moving objects by other

Figure 1: Typical frame from a bus stop video.

moving objects. A large crowd of people will occlude some individuals. It is also possible that busses and other vehicles will obscure the view of people at the bus stop, depending on the selection of camera location.

Recognition of people from a viewpoint so far away from the action is also an issue with such a system. As can be seen in the example footage in Figure 1, the resolution of this camera used in this system is not fine enough to perform accurate biometric analysis such as face recognition. Tracking of humans across the scene can also create problems. The tracker used must be able to handle following non-rigid objects. Finally, once the individuals have been recognized as such, their actions must be classified and checked for "suspiciousness."

The system is designed to use a single camera monitoring the bus stop. The system is robust in dealing with image size changes of due to perspective difference as an individual walks across the scene. Using a standard resolution of 720 by 480 pixels, the average standing person is between 80 and 130 pixels tall, depending on his or her location within the scene.

The flow chart in Figure 2 shows the layout of this system. There are three central pieces to this system: background subtraction, tracking, and human recognition.

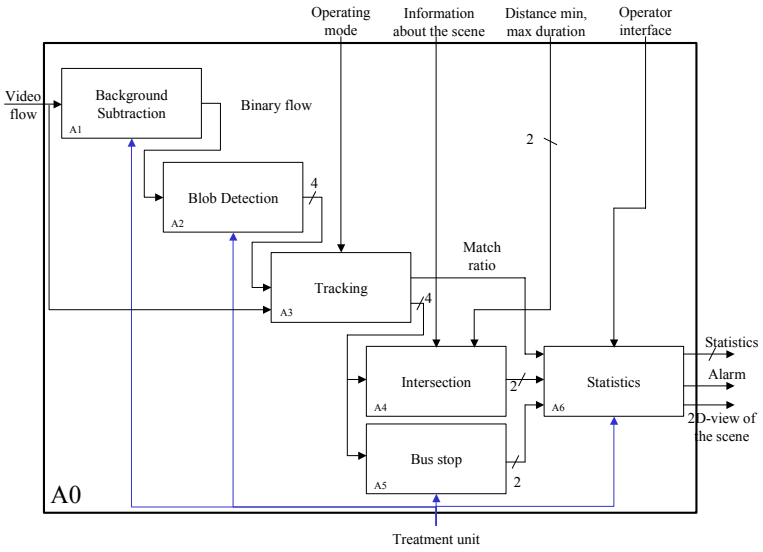

Figure 2: System flow chart.

CHAPTER 2

TRACKING OF NON-RIGID OBJECTS

Background subtraction

Background modeling is an efficient way to detect moving objects in a video sequence by comparing each new frame to this background model of the scene. In order to implement background modeling, there are simple methods such as building an average image of the scene through time, although these are not very robust. One powerful tool for building such representations is statistical modeling, where the intensity of each pixel in the video is modeled as a random variable in a feature space with an associated probability density function. Alternatively, nonparametric approaches could be used. These estimate the density function directly from the data without any assumptions about the underlying distribution. This avoids having to choose a model and estimating its distribution parameters. Elgammal *et al.* [3] have developed a general method: the kernel density estimation technique. This method is an adaptive background modeling and background subtraction technique. It is also able to detect moving objects in outdoor environments with changes in the background like moving trees or changing illumination. The implementation of the background module is based on this method.

Mean shift tracker

In many computer vision applications, such as video surveillance, it is essential to be able to track a target in real-time. Major issues with respect to tracking algorithms are partial occlusions and a moving camera. Efficiency is very important as well.

The tracking module of our system is based on a robust method by Comaniciu *et al.* [4]. This method can perform efficient tracking of non-rigid objects for which the decision process concerning the tracking is based upon the Bhattacharyya coefficient which is, in essence, a correlation score. The actual method has been simplified such that the Bhattacharyya coefficient only calculated at the end to evaluate the similarity between the target model and the chosen candidate. Thus, the method by Comaniciu *et al.* [4] is simplified into the following steps:

1. Compute the weights $\{w_i\}_{i=1\ldots n}$ according to

$$w_i = \sum_{u=1}^{m} \sqrt{\frac{q_u}{p_u(y_0)}} \delta(b(x_i) - u) \qquad (1)$$

2. Evaluate the new position y_1 according to

$$y_1 = \frac{\sum_{i=1}^{n} x_i w_i g\left(\left\| \frac{y - x_i}{h} \right\|^2 \right)}{\sum_{i=1}^{n} w_i g\left(\left\| \frac{y - x_i}{h} \right\|^2 \right)} \qquad (2)$$

where $g(x) = -k'(x)$. With the function k defined before, the expression of y_1 is much more simple:

$$y_1 = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} \qquad (3)$$

3. If $\|y_1 - y_0\| < \varepsilon$, stop the algorithm. Otherwise set $y_0 \leftarrow y_1$ and go to step 1.

The target model required for this method is characterized in our system by the color distribution in a 16-bin histogram for each RGB color channel. The number of bins for each color channel is fixed to 16 to keep the computation time down.

CHAPTER 3

HUMAN RECOGNITION

Color-based recognition

Since this system uses a single camera, individuals must be identified using a limited amount of sensory input. The field of biometrics is being researched extensively and has produced a number of methods to identify specific people. Some examples of this are fingerprint, face, and gait recognition. These are all "long-term" techniques because they are supposed to remain effective for years (e.g., a person's face takes years to change dramatically, and a fingerprint will likely never change significantly). In the scope of the bus stop monitoring project, "short-term" biometric techniques, where the measured attribute remains valid for hours rather than years, are sufficient. An example of a short-term biometric is clothing color. "The blonde man wearing a black shirt, green pants, and a purple jacket" is a description that would fit a single person at a bus stop. In our system, we use clothing color as our short-term biometric. Figure 3 shows some example snapshots of different individuals extracted from a bus stop video. We consider clothing color a very distinctive feature that should be utilized for identification.

The first step in this process is to normalize the colors in the entire scene. Assuming colors in the range [0, 1], this is done by finding the mean value for each color channel, C_k. This mean is then used to determine the correction factor for the channel that will cause the mean color to become 0.5. By normalizing the scene colors like this, the recognition module will hopefully be more resilient to slight changes in lighting.

Figure 3: Example snapshots of individuals extracted from a bus stop video.

$$C_k\,|_{k=r,g,b} = \frac{0.5}{mean(C_k)} C_k \tag{4}$$

There are different ways of quantifying clothing color. Initial tests show that using the average RGB color of a person as a database key results in many incorrect identifications. An improvement to this method segments the image of an individual into three portions based upon location within the image: head, torso, and legs. This makes intuitive sense because people typically dress in a manner that can be vertically segmented into three portions. The average color is then found for each of these regions. The vertical percentage of an image occupied by each of these three segments remains fairly constant. Using this percentage-based method is beneficial because segmentation is performed exceptionally fast. A method was attempted previously that performed the segmentation by finding the best position of two "cuts" in the image such that the total standard deviation of the pixel colors in each segment is minimized. While making intuitive sense, in practice, this method did not correctly segment the images in most cases.

Thus, each person in the database has three median colors to be compared. To recognize if two images belong to the same individual, a similarity measure is computed. The measure (d) compares the median color of the three segments as follows:

$$d = \frac{|c1_h - c2_h| + |c1_t - c2_t| + |c1_l - c2_l|}{3} \tag{5}$$

where ci_x is the median color of portion x {h:head, t:torso, l:leg} of individual i. The measure d is normalized to exist in the range [0:1]. The difference between two colors is the Euclidian distance in the RGB color space.

Drawbacks to this method include recognizing individuals who dress alike, such as a marching band, as well as people who cross into areas of deep shadows.

CHAPTER 4

RESULTS AND CONCLUSIONS

Experimental results

The system was tested on a computer equipped with a Pentium 4 2.66 GHz processor and 1 GB main memory running Microsoft Windows 2000. The tracking module works very well following people as they move across the scene. Figure 4 shows example tracking output. It can be seen that the system is successfully tracking all of the moving people in the scene. The occlusion caused by the newspaper stand and street sign in the foreground in Figure 4 is handled acceptably.

Figure 4: Example tracking module output.

Example video files showing the tracking system in action are available at http://mha.cs.umn.edu.

The tracking algorithm can be used with the system in real time. Table 1 shows results tracking a number of targets at different resolutions and the frames per second that can be used. As can be seen, tracking can be performed in real-time with color video of 320x240 resolution.

The human recognition algorithm was tested with a test set of 21 people with between three and nine images for each person (106 images total). By checking all possible combinations in this test set, the algorithm was found to have an accuracy of 82 percent. Figure 5 shows three sets of graphical images that resulted in successful matches. Also shown is the placement of

Video Color	Video Resolution	Number of Targets	Computation Speed (fps)
Color	720x480	1	25
		2	21.3
		5	12.8
		10	10.6
Color	320x240	1	>70
		5	62.5
		10	32
Grayscale	320x240	5	>70
		10	66.6
		20	62.5
		50	32.2

Table 1: Tracking module computational speed.

the two segmentation cuts. Figure 6 shows some example matches falsely determined to be the same person by the human recognition algorithm. This figure clearly illustrates the algorithm's drawbacks when multiple people dress in a similar fashion.

Conclusions and future work

In this paper, we have introduced a vision-based system to monitor for suspicious human activities at a bus stop. The system currently examines for drug dealing activity that is characterized by individuals loitering around the bus stop for a very long time without the intention of using the bus. To accomplish this goal, the system must measure for how long individuals loiter around the bus stop. To facilitate this, the system must track individuals from the video feed, identify them, and keep a record of how long they spend at the bus stop.

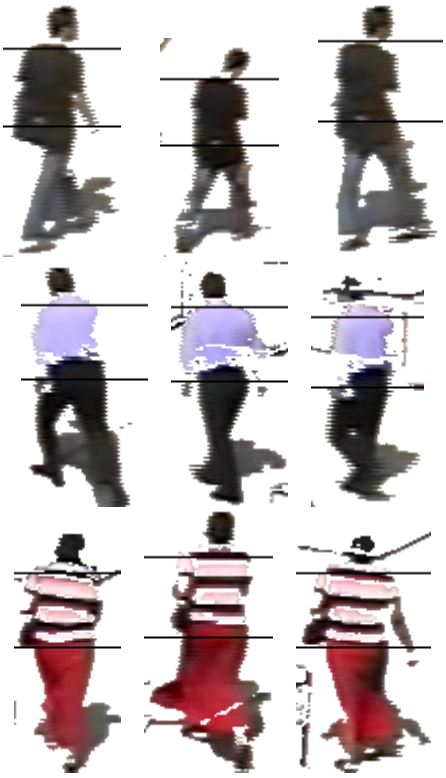

Figure 5: Three sets of successful matches for the human recognition algorithm.

The system is broken into three distinct portions: background subtraction, object tracking, and human recognition. The background subtraction and object tracking modules use off-the-shelf algorithms and are shown to work well following people as they walk around a bus stop. The human recognition module segments the image of an

Figure 6: Human recognition algorithm's false matches.

individual into three portions corresponding to the head, torso, and legs. Using the median color of each of these regions, two people can be quickly compared to see if they are the same person.

Directions for future work include improving the segmentation of body portions to better recognize individuals who have appeared in the scene previously. A method that shows great promise for segmentation is that of Efros *et al.* [11]. Using optical flow to determine which part of an image corresponds to head, torso, and legs could help improve identification of individuals by improving median color recognition for those areas.

Other completely different methods to recognize people abound as well. One possibility is to use a texture-based approach to distinguish individuals. Another possibility is to use the number of steps required to morph the image of one person into another as a heuristic to tell whether they are the same person or not.

Expanding the system further to recognize certain behaviors is also a priority. Behaviors to examine for include suspicious activities such as stretching for extended periods of time without ever jogging or leaving a package. Other actions to recognize are more benign, for instance, fainting or other medical emergencies.

Currently, the system is being re-implemented in a new framework that will allow future additions to be more quickly and easily implemented.

REFERENCES

1. C. R. Wren, A. Azarbayejani, T. Darrel and A. Pentland, "Pfinder: real-time tracking of the human body," *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, pp. 51-56, October 1997.

2. C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," *Proceedings of IEEE Computer Vision and Pattern Recognition, vol. 2*, pp. 2246-2252, June 1999.

3. A. Elgammal, R. Duraiswami, D. Harwood and L.S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proceedings of the IEEE*, vol. 90, pp. 1151-1163, July 2002.

4. D. Comaniciu, V. Ramesh and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, issue 5, pp. 564-577, May 2003.

5. N. Friedman and S. Russel, "Image segmentation in video sequences, a probabilistic approach," *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, August 1997.

6. R. Cucchiara, P. Mello and M. Piccardi, "Image analysis and rule-based reasoning for a traffic monitoring system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, n. 2, pp. 119-130, June 2000.

7. S. J. McKenna, S. Jabri, Z. Duric and H. Wechsler, "Tracking interacting people," *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 348-353, March 2000.

8. R. Rosales and S. Sclaroff, "Improved tracking of multiple humans with trajectory prediction and occlusion modeling," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Workshop on the Interpretation of Visual Motion*, 1998.

9. D. Koller, J. Weber and J. Malik, "Robust multiple car tracking with occlusion reasoning," Proceedings *of Third European Conference on Computer Vision*, vol. 1, 1994.

10. S. K. Nayar, S. A. Nene and H. Murase, "Real-time 100 object recognition system," *Proceedings of IEEE International Conference on Robotics and Automation*, vol. 3, pp. 2321-2325, April 1996.

11. A. A. Efros, A. C. Berg, G. Mori J. Malik, "Recognizing action at a distance," *Proceedings of IEEE International Conference on Computer Vision*, pp. 726-733, October 2003.