

Pedestrian Tracking from a Stationary Camera Using Active Deformable Models

Michael J. Sullivan Charles A. Richards Christopher E. Smith Osama Masoud
Nikolaos P. Papanikolopoulos*

Artificial Intelligence, Robotics, and Vision Laboratory
Department of Computer Science, University of Minnesota
4-192 EE/CS Building, 200 Union Street SE, Minneapolis, MN 55455
Phone (612) 625-0163, Fax (612) 625-0572, npapas@cs.umn.edu

Abstract

The system proposed in this paper uses active deformable models to track pedestrians moving in dynamic real-world scenes. First, figure pixels are separated from a fixed or slowly evolving ground image. Then, an initial segmentation process identifies interesting pixel blobs for tracking. The output of the segmentation process is used to choose the starting position of the control points of the active deformable model. Once tracking has begun, the control points are updated at frame rates by minimizing an energy function involving the relative position of model points, image data, and the characteristics of figure pixels.

1. Introduction

The need for pedestrian detection and tracking arises in many applications. One of these applications is the task of increasing the efficiency and safety of existing traffic systems. Providing these systems with sensory information about pedestrians crossing the streets would allow for an automatic control of traffic lights at an intersection, for example. Similarly, the dangers present in situations in which vehicles regularly travel at high speeds within meters of pedestrians, such as highway worksites, might be alleviated by the use of warning systems which can detect, locate, and track humans and vehicles. Of the sensors available, vision provides information that is richer and more complete than other sensors. In addition, because vision is the sensory modality most relied upon by humans, systems using visual input may be easier for people to install, troubleshoot, and

maintain, resulting in more reliable operation for safety-critical tasks.

Active deformable models have been used to track image gradient contours produced by objects. They offer a framework in which local image properties and some local or global model parameters can be combined, without the need for *a priori* knowledge of the scene object being tracked. Because the contour of a walking human body changes in shape and deforms continuously, we believe that methods using active deformable models to track contours are well suited for pedestrian tracking.

In this paper, we use active deformable models to track pedestrians moving in front of a static camera. The organization of the paper is as follows: Section 2 presents some previous work conducted in the area. Section 3 describes the approach we propose. In Section 4, we describe the hardware used to implement our experimental system. In Section 5, our results are discussed. Finally, in Section 6, the paper is summarized and conclusions are drawn.

2. Previous Work

Much work in traffic systems involving pedestrian control has concentrated on the control aspects [6, 7, 3]. Several researchers, however, have worked on vision-based pedestrian detection and tracking. Richards *et al.* [13] used frame differencing followed by a segmentation stage to isolate pedestrians. They enforced certain assumptions in the recognition process such as the rectangular proportions of the pedestrian body. Mori *et al.* [9] proposed a rhythm model. They extracted a signal representing the rhythmic change in shape as the pedestrian walks. This signal is

*Author to whom all correspondence should be sent.

then matched with a template to judge whether it belongs to a person. Wan *et. al.* [15] produced the TULIP system, a system for pedestrian tracking. Ali *et. al.* [1] implemented a number of detection and tracking algorithm to run on a transputer-based system. Models of the human body have been used by some researchers. Rohr [14] used a 3-D model consisting of cylinders to model the human body. Niyogi and Adelson [11] recognized pedestrians by analyzing the spatio-temporal patterns resulting from the regular motion of the legs of a walking person. The pedestrian body is then recovered and a stick model is fit to it. Other work based on human models includes [10, 12, 17]. Pedestrian recognition has also been used to automate the process of detecting border intruders, secure perimeter violators, and unauthorized safety zone trespassers [4, 8].

The concept of active deformable models, also called “snakes,” was first introduced to the field of computer vision by Kass [5]. Snakes have been used in a number of applications including image-based tracking of rigid and non-rigid objects. Using snakes requires a minimization process of an energy function. Several techniques have been used to solve this problem including variational calculus [5], dynamic programming [2], and greedy methods using heuristics [16, 18]. The latter method has the advantage of being fast as well as numerically stable. Our method uses a greedy method similar to that used by Williams and Shah [16] and Yoshimi and Allen [18]. However, we do not require objects (pedestrians in this case) to be uniformly colored as is the case in [18]. This is made possible because we use frame subtraction to isolate pedestrians from the background.

3. Approach

We propose to track the movement of pedestrians in an area observed by a stationary camera by approximating with an active deformable model the boundary of the area of image differences created by the motion of the pedestrian. As the image of the pedestrian being tracked translates and deforms in the image, continuous adjustments to the elements of the active deformable model transform and deform the boundary approximation accordingly. There are many distracting features in the difference image, such as insignificant image deformations caused by pedestrian self-motion, the appearance of other motion in the scene, and image noise. The parameters of the active deformable

model and the number of control points can be set so that these factors are ignored while following significant displacements of the pedestrian in the image plane.

3.1. Placing the Model

Movement in a scene can be detected by comparing two images acquired by a static camera: a ground image taken before the movement occurred and the current image. This difference image is defined as (where x and y are image coordinates):

$$I_{diff}(x, y) = |I_{ground}(x, y) - I_{curr}(x, y)| \quad (1)$$

To enhance the boundary contours of the pedestrian’s image in the difference image, we increase the contrast of the difference image with a simple thresholding operation, where:

$$I'_{diff}(x, y) = \begin{cases} 0 & \text{if } I_{diff}(x, y) < T \\ 255 & \text{otherwise} \end{cases} \quad (2)$$

for a threshold T . In this paper, when we use the term difference image to refer to a specific image, we mean the binary image I'_{diff} that is obtained from these operations.

When the system begins operation, a background image of the scene is captured to be used as the ground image, I_{ground} , for the calculation of image differences. When a significant motion appears in the scene, as signaled by a large region in the difference image, the image is analyzed for possible pedestrians. If the image fits the criteria established for the classification of a blob as a possible pedestrian, a bounding box is determined for the pedestrian. An active deformable model is then placed around the possible pedestrian with control points on the bounding box. Once the active deformable model has been placed, its movements are controlled by the minimization of an energy equation as described in Section 3.2.

3.2. The Active Deformable Model

The formulation of active deformable models used in this paper to approximate the pedestrian boundary draws on the work done in recent years by the computer vision community on active deformable models of contours, often referred to as “snakes.” Given a continuous contour, described as a vector:

$$\mathbf{v}(s) = (x(s), y(s)) \quad (3)$$

where s is the arc length, Kass *et. al.* [5] related the task of finding a contour in an image to the

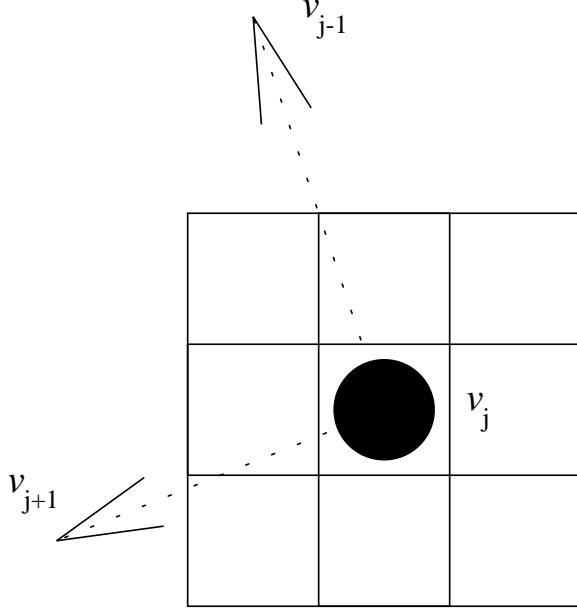


Figure 1: A single snake point in its window.

minimization of an energy function (adopting the notation used in [16]):

$$\begin{aligned}
 E_{snake}^* &= \int_0^1 E_{snake}(\mathbf{v}(s)) ds \\
 &= \int_0^1 [E_{int}(\mathbf{v}(s)) + E_{image}(\mathbf{v}(s)) + E_{con}(\mathbf{v}(s))] ds \quad (4)
 \end{aligned}$$

In this function, E_{snake}^* is the total energy of the active deformable model, E_{int} is a measure of internal energy, such as that caused by curvature, and E_{image} is a function of image characteristics. E_{con} is derived from external constraints. When this continuous model is approximated in a discrete domain (*e.g.*, a digital image) the equation becomes:

$$\begin{aligned}
 E_{snake}^* &= \sum_{j=1}^n [\alpha E_{cont}(v_j) + \beta E_{curv}(v_j) + \gamma E_{image}(v_j)] \quad (5)
 \end{aligned}$$

in which E_{cont} is derived from the distance between v_j and its neighbors, $v_{(j-1) \bmod n}$ and $v_{(j+1) \bmod n}$. E_{curv} is a function of the angle at point v_j . Again, E_{image} represents the image forces acting on the active deformable model. The terms α , β , and γ are weighting parameters which control the proportion of the active deformable

model's energy derived from each of the three terms, which are assumed to be normalized.

Kass *et. al.* [5] proposed that a minimum be found for this energy function with a variational calculus approach. Amini *et. al.* [2] have proposed a method based on dynamic programming. We have chosen to adopt the greedy method developed by Williams and Shah [16]. In the greedy method, each point on the contour is considered in turn. An energy score is calculated for locations near the current location of the control point and the control point is moved to the location which results in the lowest energy.

The E_{curv} , E_{image} , and E_{con} terms are usually sufficient to define an active deformable model approximation of an image contour when all terms vary significantly across the neighborhood of possible control point locations. However, using our current techniques, when the active deformable model is placed, it may have several control points which are far enough from the pedestrian's image that the image gradient is unvaryingly zero throughout the neighborhood of candidate locations. For these points, the term E_{image} plays no role at all and they only respond to the internal energy and external constraints, rather than to a combination of image energy and constraints.

To facilitate the initial placement of the active deformable model, we have augmented the energy equation with an E_{model} term inspired by the "balloon factor" used by Yoshimi and Allen [18] to overcome a tendency toward implosion in their active deformable models.

E_{model} is calculated as follows. First, a neighborhood of the control point in the difference image is examined. If the percentage of difference pixels set within the neighborhood falls short of a predetermined level, the control point is defined as "outside" the pedestrian's image. To bias movement of the control point toward the pedestrian's image, the locations closest to the pedestrian's image are assigned the value -1 for E_{model} . Other locations are assigned the value 0. The locations closest to the pedestrian's image can be determined because the active deformable model control points are numbered counter-clockwise around the closed active deformable model. A similar energy assignment is performed for control points which are "inside" the pedestrian's image. Besides aiding initial placement of the contour, this model energy also occasionally comes into play during later tracking stages when an object moves very quickly or has been temporarily lost for some other reason (*e.g.*, occlusion).

Most past applications of active deformable models for contour tracking have attempted to capture the entire boundary of the imaged object. Therefore, the number of control points has been chosen so that the control points fall relatively close together along the image contour. This allows the active deformable model to follow small deformations, but also makes the active deformable model vulnerable to small occlusions. One solution to this difficulty is to give the model a sense of shape through the use of internal or external constraints and to weigh these terms more strongly. This solution is not suitable for tracking pedestrians for two reasons. First, the process of tracking the pedestrian's image requires that the image forces be given considerable weight or performance degrades. Second, pedestrians do not have a well-defined shape. People must move their legs and tend to swing their arms while walking. Worse, they may turn 90 or 180 degrees, presenting an entirely different set of features and silhouette to the camera. For many purposes, it is not necessary to track these deformations. In fact, they are a distraction from the important information – the horizontal translation of the pedestrian. We have found that simply reducing the number of control points allows the model to follow a useful approximation of a pedestrian's image boundary while ignoring deformations. Because the speed of the current algorithm is linear in the number of control points, this reduction also has performance benefits.

4. The Vision Processing System

The Vision Processing System (VPS) used for these experiments is the image processing component of the Minnesota Robotic Visual Tracker (MRVT). The VPS receives input from a video source such as a camera mounted in a vehicle, a static camera, or stored imagery played back through a Silicon Graphics Indigo or a video tape recorder (see Figure 2). The output of the VPS may be displayed in a readable format or can be transferred to another system component and used as an input into a control subsystem. This flexibility offers a diversity of methods by which software can be developed and tested on our system. The main component of the VPS is a Datacube MaxTower system consisting of a Motorola MVME-147 single board computer running OS-9, a Datacube MaxVideo20 video processor, and a Datacube Max860 vector processor in a portable 7-slot VME chassis. The VPS performs the cal-

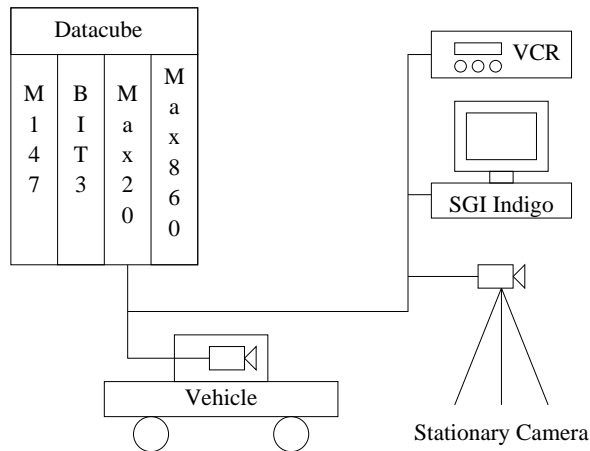


Figure 2: VPS system architecture.

ulation of the difference image and active deformable model energy minimization and calculates any desired control input. It can supply the data or the input via shared memory to an off-board processor via a Bit-3 bus extender for inclusion as an input into traffic or vehicle control software. The video processing and calculations required to produce the desired control input are performed under a pipeline programming model using Datacube's Imageflow libraries.

5. Experimental Results

The system runs at or near frame rates on the image processing system described above. At these speeds it can successfully track motion of a walking pedestrian, even when the pedestrian's image deforms in unexpected ways such as those caused by thrusting out one's arms or kicking a leg forward in an exaggerated manner. It is also fairly robust with respect to occlusions such as when two pedestrians pass in opposite directions or a single pedestrian passes behind a large tree. Potentially, more than one pedestrian could be tracked simultaneously. Although such a system should be equally robust with respect to occlusions caused by two tracked pedestrians passing one another, it would probably not be possible to tell whether the active deformable models had continued to track the same individual. Such a system might have difficulty distinguishing between two pedestrians approaching one another and then returning the way they came and two pedestrians walking past one another.

Further development of the system will require overcoming the inherent limitations of using a difference image to provide image forces for the ac-



Figure 3: A six-point active deformable model tracking a pedestrian.

tive deformable model. These problems include short and long time-scale changes in the background caused by lighting changes or continuous regular movement of objects in the scene, for example, the rustling of leaves in the wind. The system is also vulnerable to the effects of camera self-motion. A slight jitter in the camera mount could cause many patches of noise in the difference image. Although these patches will generally be ignored once contour tracking has begun, they do disturb the initial placement of the snake. Richards *et. al.* [13] describe two enhancements to the difference image framework to overcome these difficulties. First, by slowly modifying the ground image in a controlled way, changes in the background can be incorporated in the ground image. Second, to overcome the placement problem, additional processing of image regions can be done to identify portions of the image consistent with the appearance of a pedestrian. We plan to incorporate these improvements in the system described in this paper. Consideration should also be given to methods which would make it possible to mount the camera in a moving vehicle.

The current system would also benefit from a theoretical basis for the selection of the gains applied to the different elements of the energy function. Presently, these gains must be empirically determined for each application, by observing the behavior of the active deformable model in action and adjusting parameters to overcome performance deficiencies. Empirically determined gains have given satisfactory results, but a theoretical framework for gain selection would allow for the automatic determination of gains, which will be



Figure 4: The difference image which provides image forces for the active deformable model.

necessary for deployment of our system in the field by unskilled personnel.

6. Conclusion

We have presented an approach to pedestrian tracking from a static camera using active deformable models. We use these models to track the boundaries of the pedestrian's image in the difference image. By using the difference image, we avoid some difficulties associated with pedestrian tracking by tracking features, such as the occlusion of features by other parts of the pedestrian. The application of active deformable models also overcomes some of the difficulties, such as the continuous deformation of the pedestrian's image during movement, which pedestrian tracking poses to rigid model-based approaches.

Acknowledgment

This work has been supported by the National Science Foundation through Contract #IRI-9410003, the Center for Transportation Studies through Contract #USDOT/DTRS 93-G-0017-01, the Minnesota Department of Transportation through Contracts #71789-72983-169 and #71789-72447-159, the Department of Energy (Sandia National Laboratories) through Contracts #AC-3752D and #AL-3021, the 3M Corporation, the Army High Performance Computing Center, the McKnight Land-Grant Professorship Program, and the Department of Computer Science of the University of Minnesota. Michael Sullivan has also been supported by an NSF Graduate Fellowship in Visual Perception and Motor Control.

References

- [1] A. T. Ali and E. L. Dagless. Vehicle and pedestrian detection and tracking. In *Proc. of the IEE Colloquium on "Image Analysis for Transport Applications"*, page 48, 1990.
- [2] A. A. Amini, T. E. Weymouth, and R. C. Jain. Using dynamic programming for solving variational problems in vision. *PAMI*, 12(9):211–218, 1990.
- [3] P. E. An, C. J. Harris, R. Tribe, and N. Clarke. Aspects of neural networks in intelligent collision avoidance systems for Prometheus. In *Proc. of the Joint Framework for Information Technology. JFIT Technical Conference Digest*, pages 129–135, 1993.
- [4] H. Frankel, S. Riter, and A. Bernat. An automated imaging system for border control. In *Proc. of the 1986 International Carnahan Conference on Security Technology: Electronic Crime Countermeasures*, pages 169–173, 1986.
- [5] B. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1987.
- [6] F.-B. Lin. Modeling average cycle lengths and green intervals of semi-actuated signal operations with exclusive pedestrian-actuated-phase. *Transportation Research B*, 26B(3):221–240, June 1992.
- [7] S. Mathieu. Specific pedestrian crossing traffic lights. In *Proc. of the Third International Conference on Road Traffic Control*, pages 134–136, 1990.
- [8] J. C. Matter. Video motion detection for physical security applications. In *Proc. of the 1990 Winter Meeting of the American Nuclear Society*, page 396, 1990.
- [9] H. Mori and M. Sano. A guide dog robot Harunobu-5 following a person. In *Proc. of the IROS '91*, pages 397–402, 1991.
- [10] N. Murphy, N. Byrne, and K. O'Leary. Long sequence analysis of human motion using eigenvector decomposition. In *Proc. of the Intelligent Robots and Computer Vision XII: Active Vision and 3D Methods, Sponsored by SPIE*, pages 400–410, 1993.
- [11] S. A. Niyogi and E. H. Adelson. Analyzing and recognizing walking figures in XYT. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 469–74, 1994.
- [12] Y. Okawa and S. Hanatani. Determination of a pedestrian who does a prespecified body motion in a structured environment. In *Proc. of the 1991 International Conference on Industrial Electronics, Control and Instrumentation*, pages 2343–2348, 1991.
- [13] C. A. Richards, C. E. Smith, and N. P. Papanikolopoulos. Detection and tracking of traffic objects in IVHS vision sensing modalities. In *Proc. Fifth Annual Meeting of ITS America*, 1995.
- [14] K. Rohr. Towards model-based recognition of human movements in image sequences. *CVGIP: Image Understanding*, 59(1):94–115, January 1994.
- [15] C. L. Wan, K. W. Dickinson, A. Rourke, M. G. H. Bell, X. Zhang, and N. Hoose. Low-cost image analysis for transport applications. In *Proc. of the IEE Colloquium on "Image Analysis for Transport Applications"*, page 1/10, 1990.
- [16] D. J. Williams and M. Shah. A fast algorithm for active contours and curvature estimation. *CVGIP: Image Understanding*, 55(1):14–26, 1992.
- [17] G. Yan, X. Gang, and S. Tsuji. Tracking human body motion based on a stick figure model. *J. Vis. Commun. Image Represent.*, 5(1):1–9, 1994.
- [18] B. H. Yoshimi and P. K. Allen. Visual control of grasping and manipulation tasks. In *Proceedings of the 1994 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 575–582, Las Vegas, 1994.