W^4: Who? When? Where? What?
A Real Time System for Detecting and Tracking People

Ismail Haritaoglu, David Harwood and Larry S. Davis
Computer Vision Laboratory
University of Maryland
College Park, MD 20742

Abstract

W^4 is a real time visual surveillance system for detecting and tracking people and monitoring their activities in an outdoor environment. It operates on monocular grayscale video imagery, or on video imagery from an infrared camera. Unlike many of systems for tracking people, W^4 makes no use of color cues. Instead, W^4 employs a combination of shape analysis and tracking to locate people and their parts (head, hands, feet, torso) and to create models of people's appearance so that they can be tracked through interactions such as occlusions. W^4 is capable of simultaneously tracking multiple people even with occlusion. It runs at 20 Hz for 320x240 resolution images on a dual-pentium 200 PC.

1 Overview of the W^4 System

W^4 is a real time system for tracking people and their body parts in monochromatic imagery. It constructs dynamic models of people's movements to answer questions about what they are doing, and where and when they act. It constructs appearance models of the people it tracks so that it can track people (who?) through occlusion events in the imagery. In this paper we describe the computational models employed by W^4 to detect and track people and their parts. These models are designed to allow W^4 to determine types of interactions between people and objects, and to overcome the inevitable errors and ambiguities that arise in dynamic image analysis (such as instability in segmentation processes over time, splitting of objects due to coincidental alignment of objects parts with similarly colored background regions, etc.) W^4 employs a combination of shape analysis and robust techniques for tracking to detect people, and to locate and track their body parts. It builds "appearance" models of people so that they can be identified after occlusions or after other interactions during which W^4 cannot track them individually.

W^4 has been designed to work with only monochromatic video sources, either visible or infrared. While most previous work on detection and tracking of people has relied heavily on color cues, W^4 is designed for outdoor surveillance tasks, and particularly for night-time or other low light level situations. In such cases, color will not be available, and people need to be detected and tracked based on weaker appearance and motion cues. W^4 is a real time system. It currently is implemented on a dual processor Pentium PC and can process between 10-30 frames per second depending on the image resolution (typically lower for IR sensors than video sensors) and the number of people in its field of view. In the long run, W^4 will be extended with models to recognize the actions of the people it tracks. Specifically, we are interested in interactions between people and objects - e.g., people exchanging objects, leaving objects in the scene, taking objects from the scene. The descriptions of people - their global motions and the motions of their parts - developed by W^4 are designed to support such activity recognition.

W^4 currently operates on video taken from a stationary camera, and many of its image analysis algorithms would not generalize easily to images taken from a moving camera. Other ongoing research in our laboratory attempts to develop both appearance and motion cues from a moving sensor that might alert a system to the presence of people in its field of regard [9]. At this point, the surveillance system might stop and invoke a system like W^4 to verify the presence of people and recognize their actions. More generally, however, one would be interested in detecting and tracking people from a moving surveillance platform, and this is a topic currently being investigated in our laboratory also.

In W^4, foreground regions are detected in every frame by a combination of background analysis and simple low level processing of the resulting binary image. The background scene is statically modeled by the minimum and maximum intensity values and maximal temporal derivative for each pixel recorded over some period, and is updated periodically. These algorithms are described in Section 3. Each foreground region is matched to the current set of objects using a combination of shape analysis and tracking. These include simple spatial occupancy overlap tests between the predicted locations of objects and the locations of detected foreground regions, and "dynamic" template matching algorithms that correlate evolving appearance models of objects with foreground regions. Second-order motion models, which combine robust techniques for region tracking and matching of silhouette edges with recursive least square estimation, are used to predict the locations of objects in future frames. These algorithms are described in Section 4. A cardboard human model of a person in a standard upright pose is used to model the human body and to predict the location of human body parts (head, torso, hands, legs and feet). The locations of these parts are verified and refined using dynamic template matching methods. W^4 can detect and track multiple people in complicated scenes at 20 Hz speed for 320x240 resolution on 200 MHz dual pentium PC. W^4 has also been applied to infrared video imagery at 30Hz for 160x120 resolution on the same PC.

2 Previous Tracking Systems

Pfinder [1] is a real-time system for tracking a person which uses a multi-class statistical model of color and shape to segment a person from a background scene. It finds and tracks people's head and hands under a wide range of viewing condition.

[5] is a general purpose system for moving object detection and event recognition where moving objects are detected using change detection and tracked using first-order prediction and nearest neighbor matching. Events are recognized by applying predicates to a graph formed by linking corresponding objects in successive frames.

KidRooms [2, 8] is a tracking system based on "closed-world regions". These are regions of space and time in which the specific context of what is in the regions is assumed to be known. These regions are tracked in real-time domains where object motions are not smooth or rigid, and where multiple objects are interacting. Bregler uses many levels of representation based on mixture models, EM, and recursive Kalman and Markov estimation to learn and recognize human dynamics [4]. Deformable trackers that track small images of people are described in [6].

3 Background Scene Modeling and Foreground Region Detection

Frame differencing in W^4 is based on a model of background variation obtained while the scene contains no people. The background scene is modeled by representing each pixel by three values; its minimum and maximum intensity values and the maximum intensity difference between consecutive frames observed during this training period. These values are estimated over several seconds of video and are updated periodically for those parts of the scene that W^4 determines to contain no foreground objects.

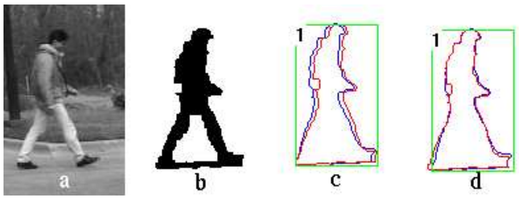

Figure 1: Motion estimation of body using Silhouette Edge Matching between two successive frame a: input image; b: detected foreground regions; c: alignment of silhouette edges based on difference in median; d: final alignment after silhouette correlation

Foreground objects are segmented from the background in each frame of the video sequence by a four stage process: thresholding, noise cleaning, morphological filtering and object detection.

Each pixel is first classified as either a background or a foreground pixel using the background model. Giving the minimum (M), maximum (N) and the largest interframe absolute difference (D) images that represent the background scene model, pixel x from image I is a foreground pixel if:

$$|M(x) - I(x)| > D(x) \quad \text{or} \quad |N(x) - I(x)| > D(x) \quad (1)$$

Thresholding alone, however, is not sufficient to obtain clear foreground regions; it results in a significant level of noise, for example, due to illumination changes. W^4 uses region-based noise cleaning to eliminate noise regions. After thresholding, one iteration of erosion is applied to foreground pixels to eliminate one-pixel thick noise. Then, a fast binary connected-component operator is applied to find the foreground regions, and small regions are eliminated. Since the remaining regions are smaller than the original ones, they should be restored to their original sizes by processes such as erosion and dilation.

Generally, finding a satisfactory combination of erosion and dilation steps is quite difficult, and no fixed combination works well, in general on our outdoor images. Instead, W^4 applies morphological operators to foreground pixels only after noise pixels are eliminated. So, W^4 reapplies background subtraction, followed by one iteration each of dilation and erosion, but only to those pixels inside the bounding boxes of the foreground regions that survived the size thresholding operation.

As the final step of foreground region detection, a binary connected component analysis is applied to the foreground pixels to assign a unique label to each foreground object. W^4 generates a set of features for each detected foreground object, including its local label, centroid, median, and bounding box.

4 Object Tracking

The goals of the object tracking stage are to:

- determine when a new object enters the system's field of view, and initialize motion models for tracking that object.
- compute the correspondence between the foreground regions detected by the background subtraction and the objects currently being tracked by W^4.
- employ tracking algorithms to estimate the position (of the torso) of each object, and update the motion model used for tracking. W^4 employs second order motion models (including a velocity and, possibly zero, acceleration terms) to model both the overall motion of a person and the motions of its parts.

W^4 has to continue to track objects even in the event that its low level detection algorithms fail to segment people as single foreground objects. This might occur because an object becomes temporarily occluded (by some fixed object in the scene), or an object splits into pieces (possibly due to a person depositing an object in the scene, or a person being partially occluded by a small object). Finally, separately tracked objects might merge into one because of interactions between people. Under these conditions, the global shape analysis and tracking algorithms generally employed by W^4 will fail, and the system, instead, relies on local correlation techniques to attempt to track parts of the interacting objects.

W^4 first matches objects to current foreground regions by finding overlap between the estimated (via the global motion model) bounding boxes of objects and the bounding boxes of foreground regions from the current frame. For each object, all current foreground regions whose bounding boxes overlap sufficiently are candidates for matching that object. Ideally, one to one matching (tracking) would be found while tracking one object. However, one to many (one tracked object splits into several foreground regions), many to one (two or more tracked objects merge into one foreground region), one to zero (disappearing) and zero to one (appearing) matchings occur frequently. W^4 tracks objects using different methods under each condition.

4.1 Appearing Objects

When a foreground region is detected whose bounding box does not sufficiently overlap any of the existing objects, it is not immediately evident whether it is a true object or a noise region. If the region can be tracked successfully through several frames, then it is added to the list of objects to be tracked.

4.2 Tracking

Here, we consider the situation that an object continues to be tracked as a single foreground region. W^4 employs a second order motion model for each object to estimate its location in subsequent frames. The prediction from this model is used to estimate a bounding box location for each object. These predicted bounding boxes are then compared to the actual bounding boxes of the detected foreground regions. Given that an object is matched to a single foreground region (and the sizes of those regions are roughly the same) W^4 has to determine the current position of the object to update its motion model. Even though the total motion of an object is relatively small between frames, the large changes in shape of the silhouette of a person in motion causes simple techniques, such as tracking the centroids of the foreground regions, to fail. Instead, W^4 uses a two stage matching strategy to update its global position estimate of an object. The initial estimate of object displacement is computed as the motion of the **median** coordinate of the object. This median coordinate is a more robust estimate of object position, and is not effected by the large motions of the extremities (which tend to influence the centroid significantly). It allows us to quickly narrow the search space for the motion of the object. However, this estimate is not accurate enough for long term tracking. Therefore, after displacing the silhouette of the object from the previous frame by the median-based estimate, we perform a binary edge correlation between the current and previous silhouette edge profiles. This correlation is computed only over a 5x3 set of displacements. Typically, the correlation is dominated by the torso and head edges, whose shape changes slowly from frame to frame. This tracking process is illustrated in figure 1.

4.3 Region splitting

An object being tracked might split into several foreground regions, either due to partial occlusion or because a person deposits an object into the scene. In this case, one object will be matched to two or more current foreground regions. W^4 determines whether the split is a true-split or a false-split (due to noise

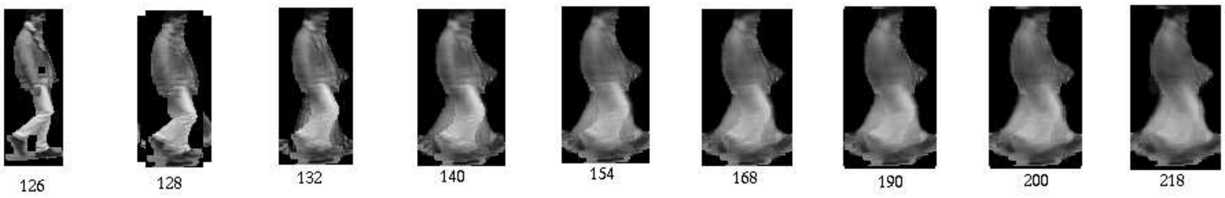

Figure 2: An example of how temporal templates are updated over time

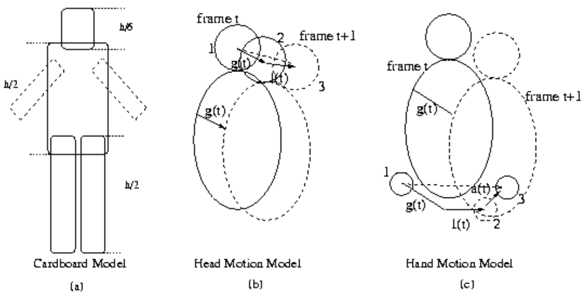

Figure 3: Cardboard model used in W^4 (a), and motion models used for the head (b) and hands (c).

transient) condition by monitoring subsequent frames, while tracking the split objects as individual objects. If W^4 can track the constituent objects over several frames, then it assumes that they are separate objects and begins to track them individually.

4.4 Region merging

When two people meet they are segmented as one foreground region by the background subtraction algorithm. W^4 recognizes that this occurs based on a simple analysis of the predicted bounding boxes of the tracked objects and the bounding box of the detected (merged) foreground region. The merged region is tracked until it splits back into its constituent objects. Since the silhouette of the merged regions tends to change shape quickly and unpredictably, W^4 uses a simple extrapolation method to construct its predictive motion model for the merged objects.

A problem that arises when the merged region splits, and the people "re-appear", is determining the correspondence between the people that were tracked before the interaction and the people that emerge from the interaction. To accomplish this, W^4 uses two types of appearance models that it constructs while it is tracking an isolated person.

W^4 constructs a dynamic template -called a *temporal texture template* - while it is tracking an isolated object. The temporal texture template for an object is defined by:

$$\Psi^t(x,y) = \frac{I(a,b) + w^{t-1}(x,y) \times \Psi^{t-1}(x,y)}{w^{t-1}(x,y)+1} \qquad (2)$$

Here, I refers to the foreground region detected during tracking of the object, and all coordinates are represented relative to the **median** of the template or foreground region. The weights in (2) are the frequency that a pixel in Ψ is detected as a foreground pixel during tracking. The initial weights $w^t(x,y)$ of Ψ are zero and are incremented each time that the corresponding location (relative to the median template coordinate)is detected as a foreground pixel in the input image. An example of how the temporal texture template of a person evolves over time is shown in figure 2.

After separation, each constituent object is matched with the separating objects by correlating their temporal templates. Since the temporal texture template is view-based, it could fail to match if there were a large change in the pose of the object during the occlusion event. Therefore, a non-view-based method, which uses a symbolic object representation, is also used to analyze the occlusion. For example, if the temporal texture templates fail to yield sufficiently high correlation values, then we match objects based on the average intensities in their upper, lower and middle parts, in an attempt to identify objects when they separates.

5 Tracking People's Parts

In addition to tracking the body as a whole, we want to locate body parts such as the head, hands, torso, legs and feet, and track them in order to understand actions. W^4 uses a combination of shape analysis and template matching to track these parts (when a person is occluded, and its shape is not easily predictable, then only template matching is used to track body parts).

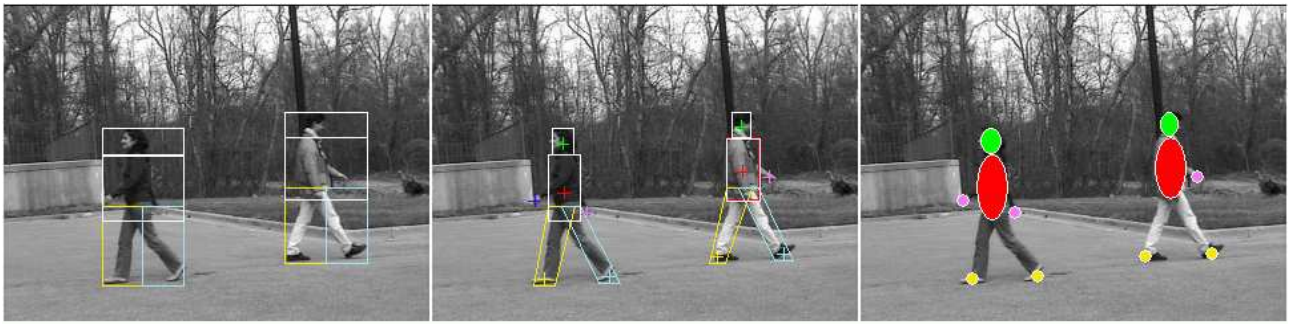

Figure 4: An example of Cardboard Model to show How Head, Torso, Legs Hands and Feet are Located. Initial Bounding boxes are located on foreground regions (a); Cardboard model analysis locates the body part (b); illustration of body part location by ellipsis (c)

The shape model is implemented using a a *Cardboard Model* [10] which represents the relative positions and sizes of the body parts. Along with second order predictive motion models of the body and its parts, the Cardboard Model can be used to predict the positions of the individual body parts from frame to frame. Figure 3 illustrates the motion models used for the hands and head. These positions are verified (and refined) using dynamic template matching based on the temporal texture templates of the observed body parts.

The cardboard model represents a person who is in an upright standing pose, as shown in figure 3(a). It is used to predict the locations of the body parts (head, torso, feet, hands, legs). The height of the bounding box of an object is taken as a height of the cardboard model. Then, fixed vertical scales are used to determine the initial approximate location (bounding box) of individual body parts, as shown in Figure 4. The lengths of the initial bounding boxes of the head, torso, and legs are calculated as 1/5, 1/2 and 1/2 of the length of bounding box of the object, respectively. The widths of the bounding boxes of the head, torso, and legs are calculated by finding the median width (horizontal line widths) inside their initial bounding boxes. In addition to finding sizes and locations, the moments of the foreground pixels inside the initial bounding boxes are calculated for estimating their principal axis. The principal axis provide information about the pose of the parts. The head is located first, followed by the torso and legs. The hands are located after the torso by finding extreme regions which are connected to the torso and are outside of the torso. The feet are located as extreme regions in the direction of the principal axes of the respective leg. Figure 4 show an example of how the cardboard model can be used to predict the locations of body parts in two stages (approximate initial location and final estimated location) and represent them as ellipsis.

After predicting the locations of the head and hands using the cardboard model, their positions are verified and refined using temporal texture templates. These temporal texture templates are then updated as described previously, unless they are located within the silhouette of the torso. In this case, the pixels corresponding to the head and hand are embedded in the larger component corresponding to the torso. This makes it difficult to accurately estimate the median position of the part, or to determine which pixels within the torso are actual part pixels. In these cases, the parts are tracked using correlation, but the templates are not updated.

6 Discussion

W^4 has been implemented in C++ and runs under the Windows NT operating system. Currently, for $320x240$ resolution gray scale images, W^4 runs at 20 Hz on a PC which has dual 200 Mhz pentium processor. For $160x120$ resolution infrared images, it runs at 30 Hz. It has the capability to track multiple people against complex background. Figure 5 illustrates some results of W^4 system in scenes of parking lots and parkland.

There are several directions that we are pursuing to improve the performance of W^4 and to extend its capabilities. Firstly, the cardboard model used to predict body pose and position is restricted to upright people. We would like to be able to recognize and track people in other generic poses, such as crawling, climbing, etc. We believe this might be accomplished based on an analysis of convex hull-like representations of the silhouettes of people. Secondly, we are investigating the integration of stereo into our detection and tracking (using the recently developed real time stereo system

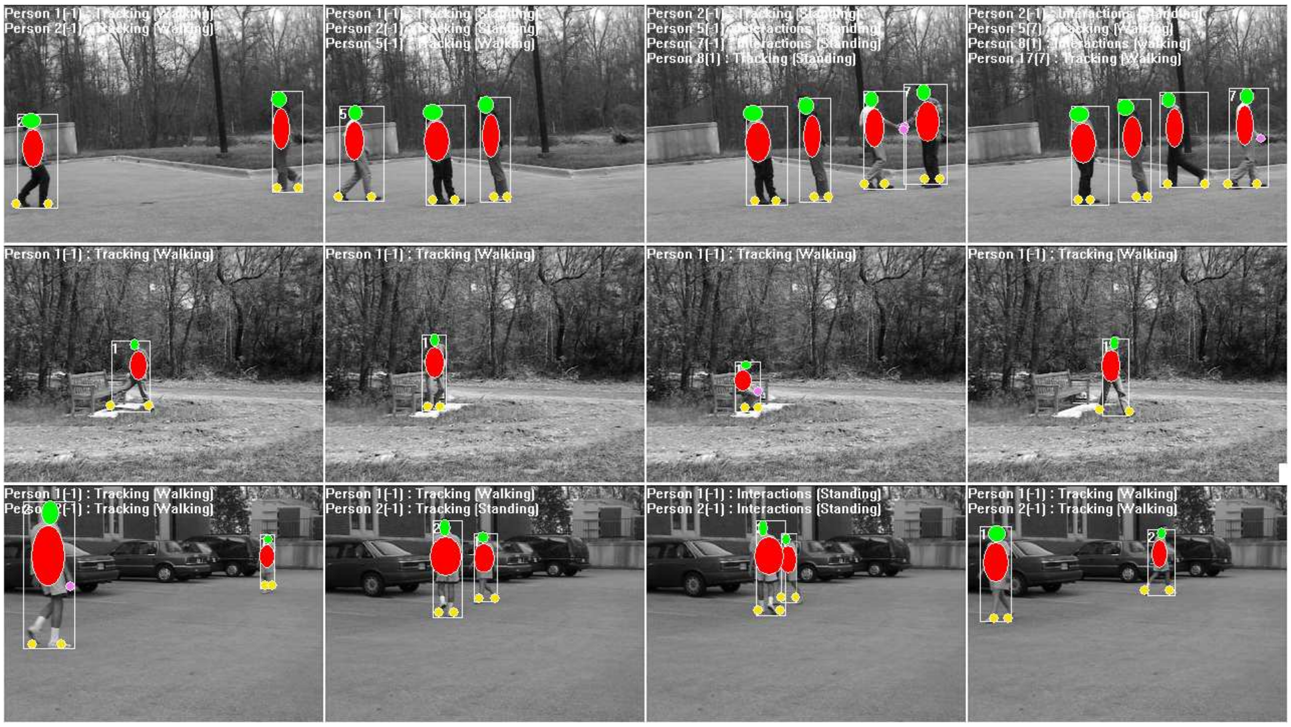

Figure 5: Examples of using the cardboard model to locate the body parts in different actions: four people meet and talk (first line), a person sits on a bench (second line), two people meet (third line).

from SRI) [11]. Stereo can be very helpful in analyzing occlusions and other interactions. Finally, our long term goal is to be able to recognize interactions between the people that W^4 is tracking. We are studying the use of temporal logic programs for the representation of actions, and to control the application of visual routines to peoples movements.

References

[1] C. Wren, A. Azarbayejani, T. Darrell, A. Pentland "Pfinder: Real-Time Tracking of the Human Body", *In Proc. of the SPIE Conference on Integration Issues in Large Commercial Media Delivery Systems*, October 1995.

[2] S. Intille, J. Davis, A. Bobick "Real-Time Closed-Word Tracking", *In Proc. of CVPR*, pp.697-703, June 1997

[3] A. F. Bobick and J. Davis "Real-Time recognition of activity using TemporalTemplates" *In Proc. Third IEEE Workshop on Application of Computer Vision*,pp.1233-1251, December, 1996

[4] C. Bregler "Learning and Recognizing Human Dynamics in Video Sequences" *In Proc. CVPR 97*, pp.569-574, June 1997

[5] T. Olson, F. Brill "Moving Object Detection and Event Recognition algorithms for Smart Cameras" *In Proc. DARPA Image Understanding Workshop*, pp.159-176, May 1997.

[6] R. Polona, R. Nelson "Low Level Recognition of Human Motion", *In Proc. Non Rigid Motion Workshop*, November 1994.

[7] A. Pentland "Machine Understanding Human Actions" *In Proc. DARPA Image Understanding Workshop*,pp.757-764, 1996.

[8] A. Bobick, J. Davis, S. Intille, F. Baird, L. Cambell, Y. Irinov, C. Pinhanez, A. Wilson "KidsRoom: Action recognition in an interactive story environment" *M.I.T. TR No: 398*, December 1996

[9] S. Fejes, L.S. Davis "Exploring Visual Motion Using Projections of Flow Fields" *In. Proc. of the DARPA Image Understanding Workshop*, pp.113-122 New Orleans, LA,1997

[10] S. Ju, M. Black, Y. Yacoob, "Cardboard People: A Parameterized Model of Articulated Image Motion", *International Conference on Face and Gesture Analysis*, 1996

[11] K. Konolige "Small Vision systems: Hardware and Implementation", *Eighth International Symposium on Robotics Research*, Hayama, Japan, November, 1997