To Appear in ACCV-98, Mumbai-India, Material Subject to ACCV Copy-Rights

Visual Surveillance of Human Activity

Larry Davis[1] Sandor Fejes[1] David Harwood[1]
Yaser Yacoob[1] Ismail Hariatoglu[1] Michael J. Black[2]

[1] Computer Vision Laboratory, University of Maryland, College Park, MD 20742
[2] Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304
lsd, fejes, harwood, yaser, hismail@umiacs.umd.edu,
black@parc.xerox.com

1 Introduction

In this paper we provide an overview of recent research conducted at the University of Maryland's Computer Vision Laboratory on problems related to surveillance of human activities. Our research is motivated by considerations of a ground-based mobile surveillance system that monitors an extended area for human activity. During motion, the surveillance system must detect other moving objects and identify them as humans, animals, vehicles. When one or more persons are detected, their movements need to be analyzed to recognize the activities that they are involved in. Ideally, the surveillance system would be able to accomplish this even while continuing to move; alternatively, the system could stop and stare at that part of the scene containing people.

In Section 1 we describe a novel approach to the problem of detecting independently moving objects from a moving ground camera, and illustrate the approach on sequences taken in very cluttered environments. Current research focuses on the problem of classifying those independently moving objects as people based on a combination of their appearance and movement. In Section 2 we describe a system that can track multiple moving people using sequences taken from a stationary camera. This system of algorithms, which has been implemented on a PC and can process 10-30 frames per second (depending on the number of people within the field of view and the resolution of the imagery) uses a hierarchy of tracking modules to identify and follow people's heads, torsos, feet, ... Finally, in Section 4 we explain how the recovered motion of these people can be classified into various activity classes using a principal component model of the time variation of motion of the body parts.

*** The support of the Defense Advanced Research Projects Agency (ARPA Order No. C635), the Office of Naval Research (grant N000149510521) is gratefully acknowledged.

2 Detecting independent motion using directional motion estimation

This section briefly describes an application of the theory developed in [2] to the problem of detecting independent motion in long image sequences. The approach, which is based on two simple geometric observations about directional components of flow fields, allows general camera motion, a large camera Field Of View (FOV), and scenes with large depth variation; no point correspondences are required. Due to the projection method the original problem of detecting independent motion is reduced to a combination of robust line fitting and one-dimensional search. More details about the method can be found in [1, 3].

2.1 Properties of directional components of visual displacement fields

Given an optical flow field, we construct a scalar field by projecting the optical flow vectors onto a given projection direction. Our approach to motion estimation is then based on analyzing cross sections of this projected flow field; in particular, cross-sections both parallel and orthogonal to the chosen projection direction. This analysis leads to recovering the projections of the camera motion, which we call the directional motion parameters. In the simple case of a narrow-FOV camera the rotational projected flow is constant along the parallel cross-sections, and varies linearly along the orthogonal cross-sections; we call this the *linearity property*. Since the projected translational flow is zero at the projection of the FOE on any parallel cross-section, the second observation leads to what we call the *divergence property*: points to the "left" of the projected FOE in a parallel cross-section have projected flow less than the flow at the projected FOE, and points to the "right" have greater projected flow. Since we do not know, a priori, what the projected rotational flow is, we estimate at *each point* along every parallel cross-section that flow value which best satisfies the divergence property. This is, essentially, equivalent to estimating a flow value that minimizes negative depth [6] along that parallel cross-section. Orthogonal cross-sections of these new projected flow values are then constructed. Finally, using the linearity property, the projected rotation parameters are estimated by finding that orthogonal cross-section on which the projected (new) flow values are best fit by a linear model. Extensions of the algorithm to large FOVs (accomplished by embedding it into a recursive derotation framework) or very small FOVs (in which the divergence property of the projected flow cannot be used) can be easily achieved. In any event, once the projected motion parameters are estimated we know by the epipolar constraint that along any parallel cross-section, all points to the left of the projected FOE should have projected motion less than the final estimate of rotation at the projected FOE obtained from the linear model, and all points to the right should have greater projected flow.

2.2 The detection algorithm

The ability to verify the epipolar constraint for arbitrary flow fields using only low-dimensional projections of the original flow field provides a simple basis for detecting independently moving objects. For this purpose we need to incorporate only one or a small collection of directional components of the flow field. The image locations where the linearity and the divergence constraints of projections are violated are considered as regions with independent motion.

Fig. 1. Detection of moving people (top and bottom) and a moving vehicle (middle) from a hand-carried camera. Each row illustrates two (non-consecutive) frames of long image sequences taken from several seconds apart.

In practice, one must take into account the fact that the linear fitting process used to estimate the projected rotation parameters must be robust to both measurement error in flow estimation and errors introduced by the presence of independently moving points; and, in the detection of independently moving points, one must take into account measurement error in flow estimation. The first problem is addressed using robust line fitting, in which the parallel cross-sections corrupted by independent motion are eliminated from the fit by a repeated-median-based robust line-estimator [7]. The second problem is addressed by first assuming that the parallel cross sections that are included in the robust line fit ("inliers") in fact do not include any independently moving points. This allows us to identify detection thresholds that adapt to changes in imaging conditions. Intuitively, for each parallel inlier cross-section we find the projected flow vector which most violates the assumption that no pixel along an inlier cross-section is moving independently. We then consider the worst violator amongst all the inlier cross-sections, and use the magnitude of the difference in flow between that pixel and the flow at the projected FOE on that cross-section as our adaptive threshold. This threshold is then applied to the remaining "outlier" cross sections. This simple automatic adaptive thresholding procedure provides a good trade-off between sensitive detection and low false alarm rate, and is a significant improvement of the detection algorithm over those which apply fixed thresholds.

In order to further improve the reliability and robustness of the algorithm, frame-by-frame-based instantaneous detections need to be integrated over both space and time. We employ temporal integration over motion trajectories using tracking to verify detections and eliminate short-term drop-outs. Finally, a spatial integration provides grouping of independently moving pixels that pass the temporal analysis based on coherence in location and velocity.

Figure 1 shows three examples of detecting independent motion from a hand-carried camera. The camera FOV is relative large (55^o) while the scenes contain different degree of depth variation. In all examples the primary input for our algorithm was the simple normal flow as a particular directional component of the flow field. In each frame dark pixels indicate local detections verified by the temporal filter. The high-lighted bounding boxes represent groupings of these detections using spatial and velocity coherence. Current research focuses on characterizing the appearance and motion of independently moving objects to classify them as people, vehicles, etc.

3 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 section we describe the computational models employed by W^4 to detect and track people and their parts. These models are designed to overcome the inevitable errors and ambiguities that arise in dynamic image analysis. These problems include instability in segmentation processes over time, splitting of objects due to coincidental alignment of objects parts with similarly colored background regions, etc.

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 [4].

W^4 consists of five computational components: background modeling, foreground object detection, motion estimation of foreground objects, object tracking and labeling, and locating and tracking human body parts. The background scene is statically modeled by the minimum and maximum intensity values and temporal derivative for each pixel recorded over some period, and is updated periodically. For each frame in the video sequence, foreground objects are detected by frame difference thresholding, connected component analysis, and morphological analysis. These foreground objects are tracked and labeled by a forward matching process from previously detected objects to currently detected objects. Motion models, which are based on matching silhouette edges of foreground objects in two successive frames and a recursive least square method, are used during object tracking to estimate the expected location of objects in future frames. A cardboard human model of a person in a standard upright pose is used to model the human body and to locate human body parts (head, torso, hands, legs and feet). Those parts are tracked using dynamic template matching methods. Figure 2 illustrates some results of the W^4 system.

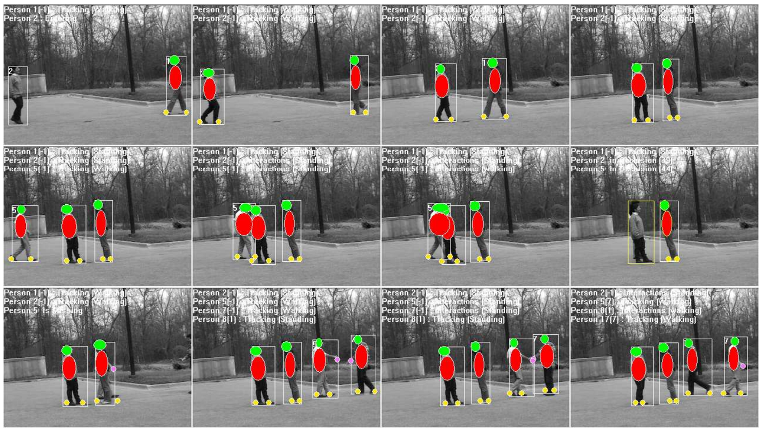

Fig. 2. Examples of using the cardboard model to locate body parts in different situations: four people meet and talk (first line), a person sits on a bench (second line), two people meet (third line).

4 Activity Modeling and Recognition

Activity representation and recognition are central to the interpretation of human movement. There are several issues that affect the development of models of activities and matching of observations to these models,

- Repeated performances of the same activity by the same human vary even when all other factors are kept unchanged.
- Similar activities are performed by different individuals in slightly different ways.
- Delineation of onset and ending of an activity can sometimes be challenging.
- Similar activities can be of different temporal durations.
- Different activities may have significantly different temporal durations.

There are also imaging issues that affect the modeling and recognition of activities

- Occlusions and self occlusions of body parts during activity performance.
- The projection of movement trajectories of body parts depend on the observation viewpoint.
- The distance between the camera and the human affect image-based measurements due to the projection of the activity.

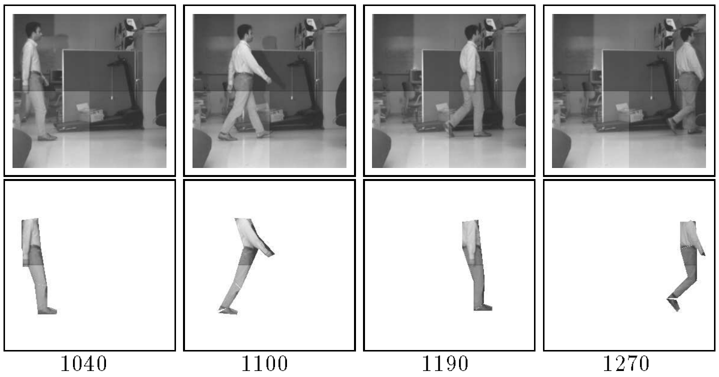

Fig. 3. Image sequence of "walking" and the five-part tracking

An observed activity can be viewed as a vector of measurements over the temporal axis. Consider as an example Figure 3, which shows both selected frames from an image sequence of a person walking in front of a camera and the model-based tracking of five body parts (i.e., arm, torso, thigh, calf and foot). We developed (see [8] for details) a method for modeling and recognition of these temporal measurements while accounting for some of the above variations in activity execution. This method is based on the hypothesis that a reduced dimensionality model of activities such as "walking" can be constructed using principal component analysis (PCA, or an eigenspace representation) of example signals ("exemplars"). Recognition of such activities is then posed as matching between principal component representation of the observed activity ("observation") to these learned models that are subjected to "activity-preserving" transformations (e.g., change of execution duration, small change in viewpoint, change of performer, etc.).

4.1 Experimental Results

We employ a recently proposed approach for tracking human motion using parameterized optical flow [5]. This approach assumes that an initial segmentation of the body into parts is given and tracks the motion of each part using a chain-like model that exploits the attachments between parts to achieve tracking of body parts in the presence of non-rigid deformations of clothing that cover the parts.

A set of 44 sequences of people walking in different directions were used for testing. The model of multi-view walking was constructed from the walking pattern of one individual while the testing involved eight subjects. The first six activity bases were used. The confusion matrix for the recognition of 44 instances of walking-directions are shown in Table 1. Each column shows the best matches for each sequence. The walkers had different paces and stylistic variations, some of which where recovered well by the affine transformation. Also, time shifts were common since only coarse temporal registration was employed prior to recognition.

Walking Direction	Parallel	Diag.	Away	Forward
Parallel	11	2		
Diagonal	3	14		1
Perp. Away			6	
Perp. Forw.	1	1	1	4
Total	15	17	7	5

Table 1. Confusion matrix for recognition of walking direction

Next, we illustrate the modeling and recognition of a set of activities that we consider challenging for recognition. We chose four activities that are overall quite close in performance: *walking, marching, line-walking*[2], and *kicking while walking*. Each cycle of these four activities lasts approximately 1.5 seconds.

We acquired tens of sequences of subjects performing these four activities as observed from a single view-point. Temporal and stylistic variabilities in the performance of these activities are common. Clothing and lighting variations

[2] A form of walking in which the two feet step on a straight line and spatially touch when both are on the ground.

also affected the accuracy of the recovery of motion measurements from these image sequences.

Table 2 shows the confusion matrix for recognition of a set of 66 test activities. These activities were performed by some of the same people who were used for model construction as well as new performers. Variations in performance were accounted for by the affine transformation. Up to 30% speed-up or slow-down as well as up to 15 frames of temporal shift were accounted for by the affine transformation used in the matching.

Activity	Walk	Line-Walk	Walk. to Kick	March
Walk	11	3		3
Line-Walk	3	24		1
Walk to Kick			12	
March	1	1		7
Total	15	28	12	11

Table 2. Confusion matrix for recognition results

References

1. S. Fejes and L.S. Davis. Detection of independent motion using directional motion estimation. Technical Report CS-TR-3815, CAR-TR-866, University of Maryland, 1997.
2. S. Fejes and L.S. Davis. Direction-selective filters for egomotion estimation. Technical Report CS-TR-3814,CAR-TR-865, University of Maryland at College Park, 1997.
3. S. Fejes and L.S. Davis. What can projections of flow fields tell us about the visual motion. To appear in *Proceedings of the International Conference on Computer Vision*, Bombay, India, January 1998.
4. S. Fejes, L.S. Davis "Exploring Visual Motion Using Projections of Flow Fields" *Proc. of the DARPA Image Understanding Workshop*, New Orleans, LA,1997
5. S. X. Ju, M. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated image motion. *Proc. Int. Conference on Face and Gesture*, Vermont, 1996, 561-567.
6. S. Negahdaripour and B.K.P. Horn. Using depth-is-positive constraint to recover translational motion. *Proceedings of the Workshop on Computer Vision*, pages 138–144, 1987.
7. A.F. Siegel. Robust regression using repeated medians. *Biometrika*, 69:242–244, 1982.
8. Y. Yacoob and M. Black Parameterized Modeling and Recognition of Activities. *To appear in ICCV-98.*