A Track-based Human Movement Analysis and Privacy Protection System Adaptive to Environmental Contexts

Sangho Park Mohan M. Trivedi
Computer Vision and Robotics Research (CVRR) Laboratory
University of California, San Diego
La Jolla, California, USA

Abstract

This paper presents a track-based system for human movement analysis and privacy protection. Our system is adaptive to environmental contexts such as illumination variations, complex moving cast shadows, different camera perspectives, and diverse site scenarios. Most of outdoor surveillance systems have been targeting at specific environmental situation: i.e., specific time, place, and activity scenarios. We address that more general human movement analysis systems should be able to handle multiple heterogeneous situations in an adaptive manner. We introduce the concept of 'spatio-temporal personal boundary' to represent different grouping patterns of human tracks, and we incorporate the concept with various site models. Experimental evaluations with extensive outdoor data show our system's robustness to environmental changes and effectiveness to properly handle various environmental contexts.

1. Introduction

Human movement analysis is important in various security and safety related applications: video surveillance, homeland/airport security, border patrol, traffic monitoring, pedestrian detection, etc. One of the main goals of such surveillance systems is to interpret what is happening in the monitored scene. Most of outdoor human monitoring systems have been targeting at specific environmental situation: i.e., specific time, place, and activity scenarios involved [3, 5, 10, 9]. We address that more general and desirable human movement analysis systems should be able to handle multiple heterogeneous situations caused by environmental variations, which requires an adaptive and robust framework. Handling multiple situations may be modeled by just adding more and more event-specific models. But the situations can be more efficiently handled by introducing a spatio-temporal structure which is common to various situations [7]. Track-based surveillance systems may be useful for various situations including different time zones such as morning, daytime, and evening time. However track data by itself does not provide much useful information about the monitored scene. In order to understand human behaviors in a scene, it is desirable to incorporate context and to recognize interrelations between multiple tracks in spatial and temporal domain. For example, similar track patterns may have very different connotation depending on different sites and contexts. Privacy protection is another important issue in video surveillance [2]. It is desirable to satisfy two seemingly contradictory goals: i.e., detailed recognition of human activity and identification blocking.

It is important to distinguish different grouping behaviors of the tracked persons in order to grasp semantic meaning of the monitored scene. Proximity-based detection of human grouping is ambiguous by itself and should depend on situational context; for example, strange people may just stand close in some situations while they are not expected to do so in other situations. In this paper we concentrate on distinguishing tracking of single persons, 2-person passing-bys, 3-person passing-bys, and grouping/splitting of persons. The concept of *personal boundary* is introduced in Section 2 to explain the group behavior of persons at different situations. Section 3 shows the system overview. Experimental evaluation and conclusion follow in Sections 4 and 5, respectively.

2. Spatio-temporal Personal Boundary

We introduce the concept of spatio-temporal *personal boundary* to explain the grouping behavior of persons. The *personal boundary* adopts the classical concept of *personal space* in social psychology [8]. The *personal space* is an invisible boundary in spatial domain into which other people may not come. If someone pierces this boundary, they will feel uncomfortable and move away to increase the distance between them. The *personal space* is adaptive in that it may enlarge or shrink depending on environmental and sociocultural contexts. For example, it shrinks in crowded areas such as shopping center and elevator, and it expands in comfort areas such as a park or a lounge. One interesting observation is that it also shrinks while people groups together in social context; i.e., people still feel secure and uninter-

0-7803-9385-6/05/$20.00 ©2005 IEEE

rupted even when other interacting persons approach closer to them. We also observe that the personal space also has proper time durations depending on the context. From the viewpoint of a single camera based 2D surveillance system, we extend the concept of personal space to define the spatio-temporal *personal boundary* as the image-based 2D boundary (i.e., around each person's bounding box but larger than the latter) that he/she wants to keep as a personal space in dynamic manner depending on situations.

The personal boundary has spatio-temporal attributes. The spatial extension of the personal boundary is specified by the same height of the original bounding box and α times width of the bounding box. The temporal duration β of the personal boundary is determined by specific duration of proximity between the spatial privacy boundaries of two persons along a sequence.

The image clips in Fig. 1 shows that two persons are waiting in front of a building entrance and the third person appears to bring them in. We understand they are forming a group in the scene, even if the three persons do not overlap. Just simple tracking functionality doe not provide enough information for the interpretation, because it is not clear whether the tracked persons form a group since the persons does not overlap or even touch one another. This example shows that we need to combine high-level context knowledge and low-level track information for a semantically meaningful surveillance. Fig. 2 depicts some of the

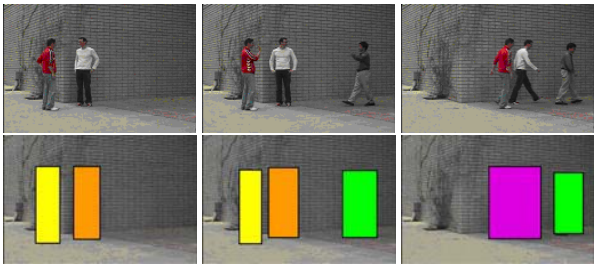

Figure 1: Detailed view of building entrance area and interacting persons; 1st row: raw input frames, 2nd row: privacy-filtered representation.

possible grouping behaviors of persons.

Fig. 3 shows the diagram of the personal boundary of two persons.

By using the *personal boundary*, we can explain different grouping behaviors of tracked persons. The adaptive factors α and β of the personal boundary may vary depending on different environmental contexts summarized in Table 1, and the factors can be trained with actual surveillance data. Table 1 shows some examples of hypothetic categorization of the spatio-temporal privacy boundaries for different times and spaces. Different sites can be modeled with different spatio/temporal adaptive factor values. The concept of spatio-temporal personal boundary may be used as a privacy protection mechanism by providing privacy grammar [2] with rich definitions of event semantics.

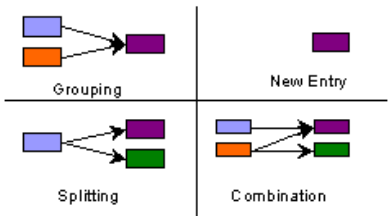

Figure 2: Multi-bounding box relations between consecutive frames: grouping, new entry, splitting, and combination.

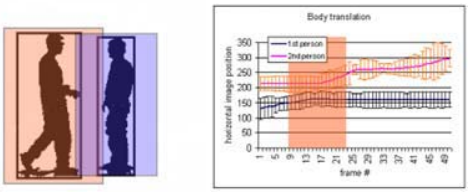

Figure 3: The illustration of spatio-temporal personal boundary. *Left*: the spatial privacy boundaries in color regions, and bounding boxes in solid straight-line rectangles. *Right*: the temporal personal boundary in brown area.

3. System Overview

Fig. 4 shows the overall system diagram. The surveillance system analyzes the input frames in sequence. Multiple background models are generated and updated online to handle background change and shadow removal. The sentry module monitors any significant change of foreground regions along the image boundary defined in terms of predefined border width, and registers the entry and exit of new persons. If no person is detected, the frame is discarded and the next frame is processed. If the same number of persons are detected as the previous frame, then the persons are updated by taking more processing steps. If the number of detected persons changes, then Expectation-Maximization (EM) learning is used to re-train the color distribution of foreground regions. The pixels are grouped into similar color blobs, and the multiple blobs are tracked. We adopt the appearance-based blob-level tracking method in [6] to associate and track multiple foreground regions. 2D Gaussian Ellipse representation is used to represent the individual humans. The tracking results are incorporated with contextual knowledge base that includes site model, track patterns, and spatio-temporal characteristics of personal boundary.

Site dependency

Site type	Crowded zone	Passage zone	Comfort zone
SB	narrow	wide	wide
TB	short/long	short	long
Examples	bus stop elevator	walkway corridor	park lounge

Activity dependency

Activity type	Pass by	Meet	Wait
SB	narrow	narrow	narrow/wide
TB	short	long	long
Examples	walkway	lounge	bus stop

Table 1: Context-dependent variations in spatio-temporal personal boundary. SB and TB denote spatial boundary and temporal boundary, respectively.

3.1. Blob Formation and Body Tracking

We adopt the codebook-based background model in [4] to segment foreground regions. Multiple codewords are generated from background training data for each pixel locations in order to represent variations in the background scene. After the background subtraction, the color distribution of foreground regions are represented by Gaussian mixture model and trained by Expectation-Maximization (EM) learning algorithm. Foreground pixels form multiple coherent blobs by region growing in attribute relational graph (ARG) according to color similarity, and the blobs are tracked between consecutive frames by ARG-based Multitarget-Multiassoiciation Tracking algorithm (ARG-MMT) [6]. Multiple blobs constitute a human body, and mis-tracked blobs are resolved at the body-level tracking. The merit of the blob-based intermediate representation of human body is that it provides efficient and versatile identity blocking mechanism for privacy protection. The blob-based privacy protection is more effective and powerful than bounding box, because it can precisely block multiple body parts independently while preserving the overall silhouette contour of the person (See 6.)

3.2. Multiple Body Tracking

The body-level tracking uses bounding boxes and 2D Gaussian representations of foreground bodies. The 2D Gaussian representation and EM-based update mechanism is effective in keeping track of grouping and splitting of people. However, the usual EM-based update is not reliable under severe occlusions or long time grouping and it can be caught in local maximum. We control the 2D Gaussian update mechanism as follows. The probabilities of classes is not updated during occlusion. We set limits for change in covariance

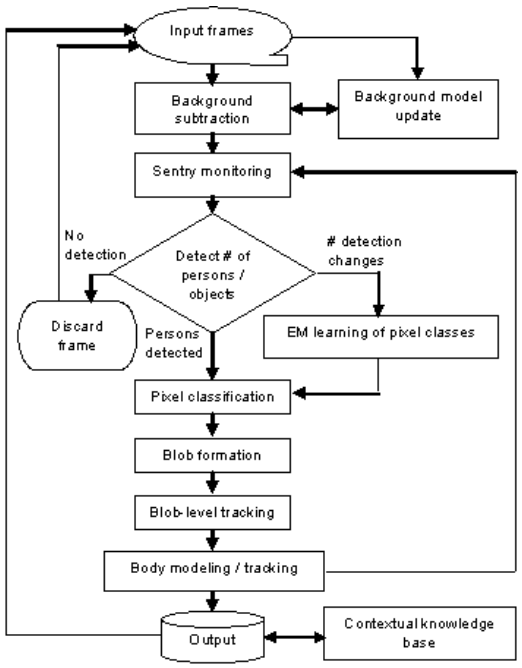

Figure 4: Block diagram of the adaptive system.

by singular value decomposition as follows: eigenvectors describe the inclination of the covariance and is only allowed little change, and eigenvalues describe the variance of the major and minor axis and is also only allowed little change. It helps avoiding classes from getting caught in local maximum by not allowing them to change considerably according to the initial human body model before occlusion. Based on tracked 2D Gaussians, assigned color blobs, and knowledge of who were present in a group before the grouping, the matching is done by calculating dissimilarity between Body models and their possible tracks with some weight factors. The weight factors weigh the individual parameters to avoid sensitivity to fluctuation in color blobs.

The body association between the tracked body T_i^{t-1} at frame $t-1$ and the new body B_j^t at frame t is performed by comparing the similarity between their blob feature vectors m_i^{t-1} and m_j^t;

$$m_i^{t-1} = [\bar{I}, \bar{J}, V_x, V_y, \mu_{hu}, \mu_{su}, \mu_{hl}, \mu_{sl}]^T for \;\; T_i \quad (1)$$

$$m_j^t = [\bar{I}, \bar{J}, V_x, V_y, \mu_{hu}, \mu_{su}, \mu_{hl}, \mu_{sl}]^T for \;\; B_j \quad (2)$$

where $\bar{I}$, $\bar{J}$ are the mean position of the blob and V_x,V_y are the mean motion vector in horizontal and vertical image dimensions obtained by block-matching based motion estimation. μ_h, μ_s are the mean intensities of H and S channels in HSV color space. The subscripts (u) and (l) indicate the upper and lower body regions, respectively. Π_{t-1} and Π_t are the covariance matrices of these features for all the

tracks in the image at time $t-1$ and all the blobs at time t, respectively.

We assign weights for each of the feature components with

$$W = [w_x, w_y, w_{vx}, w_{vy}, w_{hu}, w_{su}, w_{hl}, w_{sl}], \mid W \mid = 1 \tag{3}$$

The weighted Mahalanobis distance $\Delta_{ij}^{t-1,t}$ uses the weighted dissimilarity D_w between the i-th track T_i^{t-1} at time $t-1$ and the j-th blob B_j^t at time t as follows;

$$\Delta_{ij}^{t-1,t} = (D_w)^T (\Pi_{t-1} + \Pi_t)^{-1} (D_w) \tag{4}$$

$$D_w = W. * m_i^{t-1} - W. * m_j^t \tag{5}$$

where $W. * m$ denotes the component-wise multiplication between weight vector W and feature vector m, producing a weighted vector D. In the actual implementation, the covariance matrices Π_{t-1} and Π_t are assumed to be diagonal, simplifying the computation of $\Delta_{ij}^{t-1,t}$.

Figure 5: Track analysis of 'bypassing' sequence. Bypassing is detected and marked as green rectangle.

4. Experimental Studies

NTSC videos of pedestrians are captured at 30 frames per seconds speed from various outdoor environments such as building entrance, walkway, bus station, etc. The video data were captured from various perspectives including camera inclinations of 0, 25, 35, and 55 degrees from horizontal level. Some videos were captured with Genwac GW-202D CCD camera, and others with Sony HandyCam DCR-TRV950.

We performed three kinds of tests. Test 1 is to estimate the tracking performance at different camera perspectives, and involves building entrance and walkway views captured from ground 1st, 2nd and 3rd floor at 2 o'clock in the afternoon under cloudy weather condition. Fig. 6 shows some example frames of walkway captured from the 2nd and the 3rd floor. Table 2 shows the input data summary for Test 1. We subsampled the video sequences that include pedestrians whose actions were choreographed. The subsampled videos corresponds to 30 minutes long sequence. The analysis of detection errors and tracking errors are shown in Table 3. False alarm denotes the cases that the tracker detects noise foreground blobs as pedestrians. Missed detection (partial) represents the cases that some body regions

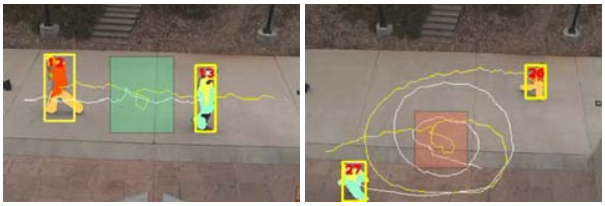

Figure 6: Multi-person tracking from various perspectives. Bypassing as green and grouping as red rectangles.

Input data summary

Input data	Count
Number of persons	207
Number of frames	56603
Event	**Count**
Single person	23
2-person passing-by	53
3-person passing-by	16
Entry as a group	11

Table 2: Test 1 data summary

are not detected. Missed detection (complete) represents the case that a pedestrian is not detected. Redundant detection denotes a single person is detected as multiple persons. Track switch (temporary) means that some tracks are incorrectly switched from each other but recovered. Track switch (permanent) means that two tracks are switched and person identity gets confused. Track lost means that a track is lost during tracking. Our surveillance system is robust to various perspectives in representing and tracking human bodies.

Test 2 is to evaluate the robustness of the surveillance system with a very long sequence of outdoor video data. We captured a walkway scene from 9 AM to 7 PM, resulting in

Detection errors

Type of error	Count	[%]
False alarm	5	
Missed detection (partial)	10	5
Missed detection (complete)	1	0.5
Redundant detection	6	3

Tracking errors

Type of error	Count	[%]
Track switch (temporary)	13	19
Track switch (permanent)	21	30
Track lost	1	0.5

Table 3: Test 1 error analysis

about 10 hour video composed of over one million frames. The all-day long outdoor video sequence contains various kinds of environmental changes. Fig. 7 shows some examples of varying illumination conditions of the same site from the morning, noon, afternoon, and evening time.

Figure 7: Illumination change of a walkway along a day. The capture times are approximately 9 AM, 1 PM, and 7 PM from left to right, respectively.

The long video sequence involves dramatic variations in average illumination level, moving shadows from wind-blown branches, drastic changes of intensity histogram profile, etc. Fig. 8 shows 3D view of concatenated histogram profiles along a day, and Fig. 9 shows some instances of the histogram profiles and the variation of average illumination levels of each frames.

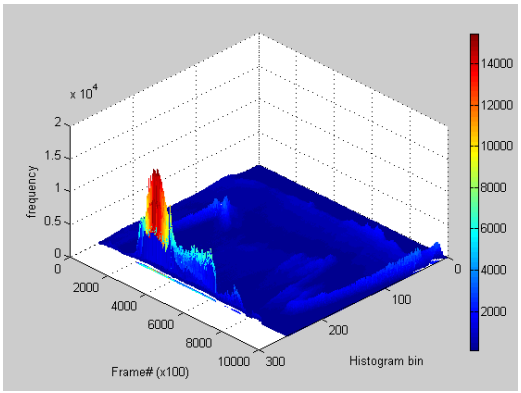

Figure 8: Histogram profile of the walkway along a day.

The long video sequence is a natural scene and the persons captured in the scene are anonymous pedestrians; no artificial treatment was made toward the pedestrians. The majority of the video frames are background scenes without any pedestrians; therefore we arbitrarily subsampled the video clips that contain pedestrians. Table 4 shows the input data summary for Test 2.

Some example frames of multi-person and single-person tracking results are shown in Fig. 10 including noon, afternoon, and evening time. The analysis of detection errors and tracking errors are shown in Table 5. Most of the false alarms are due to background subtraction noise.

Test 3 is to apply our context-adaptive surveillance system to monitor pedestrian safety at bus stations. We present

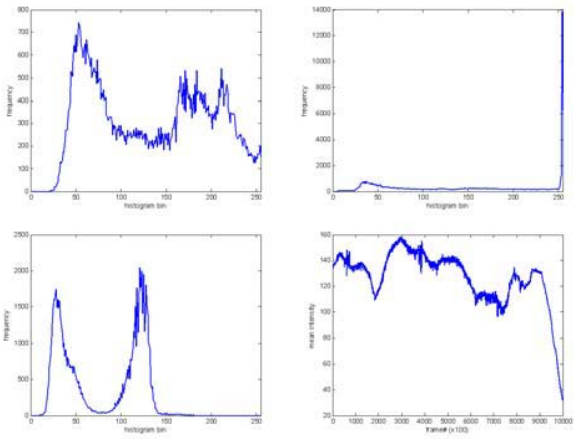

Figure 9: Example histogram profiles of morning, noon, and evening time, and the mean intensity variation along a day at the walkway.

Input data summary

Input data	Count
Number of persons	15
Number of frames	21600
Event	**Count**
Single person	9
2-person passing-by	3
3-person passing-by	0
Entry as a group	1

Table 4: Test 2 data summary

some of the results. Multiple background models are trained with video frames that contains no cars or stationary cars as shown in Fig. 11. The spatial structure of the site [1], depicted in the right image of Fig. 11, is stored in a surrogate spatial database. The watch zones such as driving area (in red) and pedestrian areas (in green) are manually assigned in terms of non-overlapping region of interest (ROI).

The surveillance system detects, stores, and queries this spatial database and significant events with their spatial and temporal attributes. The left image of Fig. 12 shows tracked pedestrians and their grouping event (in greed rectangle.) The middle and right images of Fig. 12 show the detection

Figure 10: Multi-person tracking examples.

Detection errors

Type of error	Count
False alarm	11
Missed detection (partial)	1
Missed detection (complete)	0

Tracking errors

Type of error	Count
Track switch (temporary)	0
Track switch (permanent)	1
Track lost	0
Track drift onto other person	2

Table 5: Test 2 error analysis

Figure 11: Example frames contained in multiple backgrounds for bus station and site modeling; red and green regions represent watch zone and pedestrian zones, respectively.

and tracking of pedestrians and a moving car in the watch zones, respectively.

Figure 12: Bus station tracking results.

5. Conclusion

In this paper, we have presented a track-based surveillance and privacy protection system that is adaptive to outdoor environmental contexts. The environmental contexts involve varying illuminations, changing weather conditions, complicated moving cast shadows, various camera perspectives, and site variations depending on locations. We have introduced the spatio-temporal *personal boundary* in order to address the different grouping behaviors of persons in different environmental contexts such as walkway and bus station. The experimental tests show that the system is robust to environmental fluctuations and effective to handle various site scenarios. We also presented a promising example result about the application of the system to pedestrian safety monitoring at a bus station.

Acknowledgments

This research is supported in part by the Technical Support Working Group (TSWG) of the US Department of Defense, by a University of California Discovery Grant under the Digital Media Innovation Program, and by NSF RESCUE ITR Project. We are also thankful for the assistance by the visiting students from Aalburg University in Denmark, Rasmus Corlin and Preben Andersen, and for the support of our colleagues from the UCSD Computer Vision and Robotics Research Laboratory.

References

[1] S. Bhonsle, M. M. Trivedi, and A. Gupta. Database-centered architecture for traffic incident detection, management, and analysis. In *IEEE Conference on Intelligent Transportation Systems*, Dearborn, Michigan, October 2000.

[2] D. Fidaleo, H. Nguyen, and M. Trivedi. The networked sensor tapestry: A privacy enhanced software architecture for interactive analysis and sensor networks. In *ACM 2nd International Workshop on Video Surveillance and Sensor Networks*, 2004.

[3] I. Haritaoglu, D. Harwood, and L. Davis. Hydra: Multiple people detection and tracking using silhouettes. In *Proceedings of Second IEEE Workshop on Visual Surveillance*, pages 6–13, Fort Collins, USA, 1999.

[4] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis. Background modeling and subtraction by codebook construction. In *IEEE International Conference on Image Processing*, 2004.

[5] S. J. McKenna, S. Jabri, Z. Duric, and H. Wechsler. Tracking interacting people. In *4th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2000)*, pages 348–353, 2000.

[6] S. Park and J. K. Aggarwal. Segmentation and tracking of interacting human body parts under occlusion and shadowing. In *IEEE Workshop on Motion and Video Computing*, pages 105–111, Orlando, FL, 2002.

[7] S. Park and J.K. Aggarwal. A hierarchical bayesian network for event recognition of human actions and interactions. *Multimedia Systems: Special Issue on Video Surveillance*, pages 164–179, 2004.

[8] R. Sommer. *Personal Space: The Behavioral Basis of Design*. Prentice Hall, Englewood Cliffs, New Jersey, 1969.

[9] M. M. Trivedi, T. Gandhi, and K. Huang. Distributed interactive video arrays for event capture and enhanced situational awareness. *IEEE Intelligent Systems, Special Issue on Artificial Intelligence for Homeland Security*, September 2005.

[10] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9), 2004.