Observation-Switching Linear Dynamic Systems for Tracking Humans Through Unexpected Partial Occlusions by Scene Objects

Patrick Peursum Svetha Venkatesh Geoff West
Dept of Computing, Curtin University of Technology GPO Box U1987, Perth, Western Australia
{peursump, svetha, geoff}@cs.curtin.edu.au

Abstract

This paper focuses on the problem of tracking people through occlusions by scene objects. Rather than relying on models of the scene to predict when occlusions will occur as other researchers have done, this paper proposes a linear dynamic system that switches between two alternatives of the position measurement in order to handle occlusions as they occur. The filter automatically switches between a foot-based measure of position (assuming $z = 0$) to a head-based position measure (given the person's height) when an occlusion of the person's lower body occurs. No knowledge of the scene or its occluding objects is used. Unlike similar research [2, 14], the approach does not assume a fixed height for people and so is able to track humans through occlusions even when they change height during the occlusion. The approach is evaluated on three furnished scenes containing tables, chairs, desks and partitions. Occlusions range from occlusions of legs, occlusions whilst being seated and near-total occlusions where only the person's head is visible. Results show that the approach provides a significant reduction in false-positive tracks in a multi-camera environment, and more than halves the number of lost tracks in single monocular camera views.

1 Introduction

Recent research in human trajectory tracking has focused on handling scenarios where multiple humans mutually occlude each other as they move about the scene. The fact that the occluding and occluded objects are both tracked humans is often used to infer when an occlusion is occurring (eg: [5, 8, 13, 14]). However, it is not possible to use similar reasoning when occlusions are caused by unmodelled, stationary scene objects. To address this, a switching linear dynamic system [6] is proposed that switches on the *observations* (rather than the state) to improve the robustness of tracking a human through frequent partial occlusions by scene objects. No *a priori* information on the scene or its contents is required. The filter tracks through occlusions of a person's lower body by switching between two alternative measurements of the person's position, one based on their feet and the other on their head and estimated height. The premise is that the two alternative measurements will be suitable for different situations — head-based estimates are more accurate during times of occlusion and foot-based estimates are more accurate during times of unobstructed view. The concept of alternative position measurements is not specific to the linear dynamic system, hence it is applicable to a variety of filters such as the Extended Kalman filter, multi-hypothesis (dynamics-switching) filters or joint probability data-association filters (JPDAFs). To demonstrate this, two observation-switching linear dynamic systems (O-SLDSs) are evaluated — one based on a Kalman filter (O-SKF) and another on a particle filter (O-SPF). Both are compared against their unswitched counterparts.

In this paper, a person's foreground silhouette is extracted via background subtraction [11]. Tracking occurs by maintaining two independent but interacting filters for each person — an O-SLDS for the person's 3D position and a standard (unswitched) Kalman or particle filter for the person's 3D height. The standard filter on height is used to calculate the head measurement of the person's position and facilitates tracking even when the person intentionally changes height (such as by sitting down). This is in contrast to other research where position is calculated exclusively via the head measurement [14] by assuming that height is pre-defined and unchanging, or where fixed human-sized rectangles are matched against the observed silhouettes [2]. In both cases, humans are assumed to be close to a pre-defined 'average' height in order for position estimation to be accurate, precluding the possibility that a person changes height (eg: by kneeling or sitting down).

This paper demonstrates that more than one alternative of the same measurement (in this case, position) can significantly reduce errors in tracking in comparison to a standard filter. The utility of the O-SKF and O-SPF is evaluated by tracking a person in a furnished room through various levels of occlusions in one or more camera view of the scene, including occlusions of all but the person's head as well as occlusions whilst being seated. The signif-

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006

icance of the approach is summarised as follows. (1) No prior knowledge of occluding objects in the scene is required. (2) Tracking can handle cases where the person changes height while occluded, such as when sitting down during an occlusion. (3) The approach improves robustness in both single monocular views and multi-camera environments. (4) The computational complexity of the O-SLDS can be reduced to that of a standard unswitched filter by noting that the switch variable (modelling occlusion) is effectively observable based on the relative accuracy of the two measurements.

2 Related Work

With monocular cameras, the world position of a person is often estimated by assuming that the person is moving along the ground-plane, $z = 0$ (eg: [2, 11, 14]). If the camera is calibrated to the world coordinate system, the person's position can be estimated by (1) recovering their silhouette (eg: via background subtraction), (2) assuming the lowest point of the silhouette corresponds to the person's feet, and (3) mapping this point into 3D by fixing $z = 0$. This approach is often used even if multiple cameras monitor the scene since stereo correspondence is unreliable when working with wide-baseline views due to the substantial changes in viewpoint and object appearance, particularly if occlusions are present.

Since the feet of a person can easily become occluded, some researchers [14] instead map the person's head point into 3D by assuming that their height is close to an 'average' height H_{avg} and fixing $z = H_{avg}$ during the 2D to 3D mapping. Recent work by Fleuret *et al* [2] proposed a radically different approach to estimating the location of people in a crowded room. They search for the maximum-likelihood set of discretely-positioned, human-sized rectangles that best explain the observed silhouettes. This approach handles both complex person-person occlusions and occlusions by scene objects, but tracked people must still be standing upright in order to reasonably fit the fixed-height, human-sized rectangles. Thus common actions such as sitting down will invalidate the system's height assumption.

Regardless of how the 3D position is estimated, tracking filters generally expect that errors in position measurements are due to random noise, typically assumed to be zero-mean Gaussian. Observed positions are then assigned ('gated') to tracking filters either deterministically (nearest-neighbour matching or similar) or via a 'soft-gated' (probabilistic) approach such as the JPDAF [1]. However, in the case of an occlusion of the legs, the measurement error of the foot-based position is a (linear) function of the occlusion's unknown extent *plus* the measurement noise. This will cause large errors in position and result in gating failures. Using a head-based measurement of position avoids problems from occlusion but introduces a similar issue when the person changes height (eg: sits down) since the measurement error will now be a (linear) function of the difference between the unknown true height and the assumed average height, plus noise, and gating will again fail.

Some researchers have explicitly considered the effect of occlusions by scene objects on tracking. As part of a range of methods aimed at improving the robustness of tracking humans, Xu and Ellis [12] manually define the areas of a scene that may occlude people (termed 'static occlusions'). They then use this foreknowledge of occluding areas to assist in tracking. Similarly, a depth map of the scene derived from human motion [3, 4] can be used to predict occlusions. Indeed, in [3] the authors use a depth map to assist in appearance-based tracking by consulting the depth map to determine what portion of the human should be visible at every time frame as they walk through the scene. However, the depth map must be pre-learned before it can assist in tracking, and must be re-learned when the scene changes (eg: if an occluding object is removed or introduced).

In many indoor scenes such as households and offices, the placement of furniture is liable to change over time. Hence this paper suggests that the tracking filter itself should handle unexpected occlusions by switching to an alternative measurement of the person's location during an occlusion. Switching on observations has been employed before, but as meta-data describing the configuration of the observations [7, 10] rather than as a means of alternating between two versions of the same observation.

3 The O-SLDS for Tracking Through Partial Occlusions

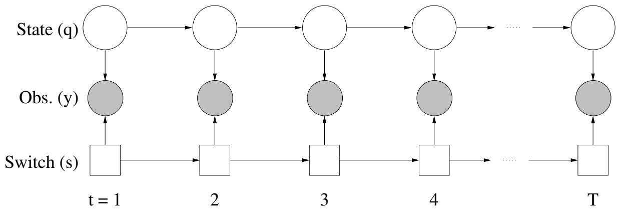


Figure 1. A switching linear dynamic system switched on the observations. Observed nodes are shaded. Circles represent continuously-valued nodes, squares indicate discretely-valued nodes.

Figure 1 shows the general form (as a dynamic Bayesian network) of an O-SLDS [6]. In this paper, the observation y_t comprises of a tuple $\{y_t^H \; y_t^F\}$, representing the position measurements based on the head and feet respectively. The switch s_t models the need to 'select' one of these observation alternatives over the other. In this paper, s_t is a Boolean variable representing whether or not an occlusion is occurring (ie: $s_t \in \{Occ, NoOcc\}$).

3.1 Improving the Efficiency of the O-SLDS

In a system where observations are allocated to tracks via deterministic-gating, the computational complexity of the O-SLDS can be reduced by analytically determining the state of s_t.[1] This can be achieved by modifying the gating procedure to take into account the relative accuracy of the head- and foot-based position measurements y_t^H and y_t^F respectively. These two measurements are calculated by:

$$y_t^H = \phi\big((i_x^H, i_y^H), w_z = Height_{t|t-1}\big) \quad \text{(1a)}$$

$$y_t^F = \phi\big((i_x^F, i_y^F), w_z = 0\big) \quad \text{(1b)}$$

where:

- ϕ is retrieved from camera calibration and is the transformation from the image coordinates (i_x, i_y) to world coordinates (w_x, w_y, w_z) when given w_z;
- (i_x^H, i_y^H) are the image coordinates of the centre-top point of the observed blob's bounding box;
- (i_x^F, i_y^F) are the image coordinates of the centre-bottom point of the observed blob's bounding box; and
- $Height_{t|t-1}$ is the predicted world height of the blob at time t given evidence up to time $t-1$, extracted from a second filter on height (see Section 3.2).

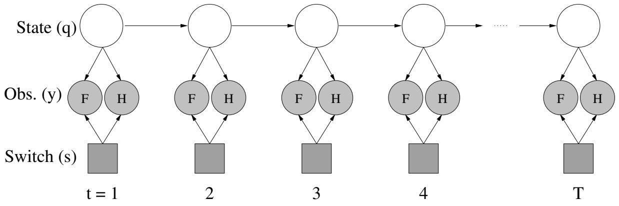


Figure 2. The O-SLDS used in this paper. The two components of the y_t tuple (F and H) are explicitly depicted here. s_t is observable and assumed independent of s_{t-1}.

Figure 2 shows the O-SLDS used in this paper. The two alternative measurements are assumed to be independent of each other given the state and switch value, and the latter is determined as follows. During an occlusion of a person's lower body, y_t^F will not relate to the true position of the person since the assumption that $w_z = 0$ at y_t^F is no longer valid. However, if $Height_{t|t-1}$ is accurate, y_t^H will still be close to the true position and, by extension, close to the position predicted by the tracking filter. Given this, gating is performed by pairing together the two alternative measurements of each blob. Tracks are then allocated measurement pairs in a nearest-neighbour manner using the Mahalanobis distance between the predicted and the observed positions. For allocation purposes, the individual measurements in the pairs (y_t^H or y_t^F) are treated separately — the pair which contains any measurement (head or foot) that minimises the distance to a track is assigned to that track. The value of s_t is then set based on which measure in the pair is closest to the predicted position. If y_t^H is more accurate, $s_t = Occ$; conversely, a more accurate y_t^F implies $s_t = NoOcc$.

Since s_t is analytically observable, three assumptions may be made to reduce the inference complexity and number of parameters of the switching Kalman filter. The first is to assume that since s_t is now fully observable, it can be made independent of s_{t-1}. This also alleviates the need to estimate the transition parameter $P(s_t|s_{t-1})$. For similar reasons, a second simplification is to assume that the prior probability of occlusion is uniform (ie: $P(s_t = Occ) = P(s_t = NoOcc) = 0.5$). The final simplification is to fix the probability of the 'wrong' measurement in the pair to 1.0. That is, $P(y_t^F|s_t = Occ) = 1.0$ and $P(y_t^H|s_t = NoOcc) = 1.0$. This assumption is justified by noting that when an occlusion is occurring, y_t^F will be some (unknown) linear translation of the true measurement due to the occlusion. Hence y_t^F should be considered unobservable during an occlusion and so must be marginalised out. Similarly, when no occlusion is occurring, y_t^H is less accurate than y_t^F and so is marginalised out.

Taken together, these three assumptions reduce the complexity of the O-SLDS to that of a standard (unswitched) linear dynamic system. Specifically, the wrong measurement is always marginalised out at every time t. Since it has no effect on the joint probability (ie: probability = 1.0) and since the prior on the switch is uniform, the wrong observation at each time t can simply be omitted from the network. Computationally, this involves providing the correct measurement as the *only* observation for the filter and discarding the wrong measurement. In this way, the computations that are *necessary* for the O-SLDS will reduce to a complexity equal to an unswitched filter, with *unnecessary* computations eliminated by modifying the process of observation gating and measurement selection.

3.2 The Role of Height Estimation

In Equation (1a), the value of $Height_{t|t-1}$ is critical in correctly estimating the position of a person based on their head location. $Height_{t|t-1}$ is retrieved from predictions made by a second unswitched filter that tracks the person's world height. After blobs are assigned to tracks using the O-SLDS, height is estimated as per [9] and the height filter is updated — it plays no part in gating. If a blob is occluded, its position will have been calculated from y_t^H and so its height will equal $Height_{t|t-1}$. If all views are occluded, $Height$ will thus become a constant, fixed to the 'last known good' height. In effect, this worst-case situation is equivalent to the fixed-height assumption of [14].

[1] This could also be done for soft-gated approaches such as the JPDAF, but must be performed over all possible combinations of observations and tracks. Furthermore, it makes little sense to avoid stochastic selection of measurements when tracks are already stochastically gated.

4 Experimental Setup

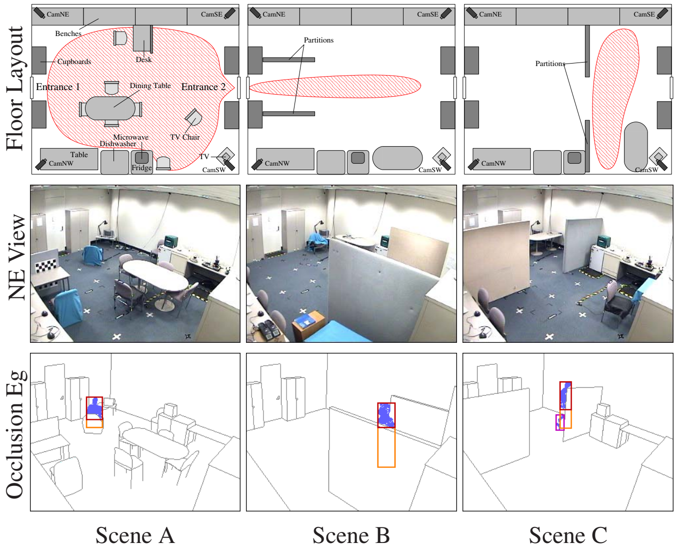


Figure 3. The three scenes used to evaluate the O-SKF and O-SPF. Greyscale intensity in the floor layouts indicate height of each object (darker = taller). The hashed region (red) indicates the area where the person moves in each scene. The lowest row shows head-based tracking during occlusions (scene outlines are for the reader's reference).

Experiments were conducted capturing a person moving through scenes containing tables, chairs and partitions. The scene is monitored by four fixed cameras, one in each corner of the room capturing video at a resolution of 320×240 pixels, 25 frames per second. Background subtraction is used to extract a moving person's foreground silhouette.

Three scenes were arranged that contained different types and placements of occluding obstacles. Figure 3 shows the floor layout, north-east camera view and examples of occlusion in each scene. Scene A involves the person walking around a furnished room for several minutes, generating partial occlusions in one or more camera views for over 80% of all video frames. This includes occlusions during changes in height, such as when the person sits down in one of the four chairs in the scene. Scene B has the room configured such that two 1.8m-tall partitions flank the entrance, occluding all but the head of the person in two camera views as the person enters the room. The person repeatedly enters and leaves the room. Scene C has the same partitions arranged in a 'walled-off' manner with a gap existing between the two partitions. The person walks back and forth parallel to the partitions. At least one of the NW and NE views is occluded in over 90% of frames — only the person's head is visible in these views during an occlusion.

Each scene is processed using four different filters — two standard filters (a Kalman filter and a particle filter) and two observation-switching filters (an O-SKF and an O-SPF). The particle filters both use 1000 particles to approximate the positional distribution. All filters are based on constant-velocity dynamics with identical model uncertainties — 60mm and 30mm/frame for position and velocity standard deviations respectively.

5 Results and Analysis

5.1 Multiple Camera Views

Figure 4 plots the number of tracked objects over all frames of each of the three scenes, comparing the detection of false-positives for each of the four filters. Table 1 provides a numerical evaluation of the same results. Only one person moves in the scene. False-positives occur when the person is partially occluded in a view — the person will appear to be several metres further away from the camera then they actually are (according to the foot measurement y_t^F). Multiple occluded views thus lead to multiple false-positive tracks, one per occluded view. Since at least one camera always has an unobstructed view of the object in this paper, object counts over one indicates false-positive tracks.

It can be seen that the O-SKF and O-SPF significantly reduce the number of false-positive tracks in comparison to the standard Kalman and particle filters. Note that not all false-positives are eliminated by the two O-SLDS filters. This is for a variety of reasons:

- *Background Subtraction Errors*. These produce false-positive blobs, such as the breaking up of a single person into several disjoint foreground blobs, and false blobs generated by noise, strong shadows and reflections. Under-segmentation of the person's silhouette can also cause inaccuracies in $\{y_t^H, y_t^F\}$.
- *Noisy Occlusions*. Occlusions are not always 'clean'. Gaps in occluding objects break up foreground silhouettes and a person's feet often trail or lead the person's body on either side of an occluding object
- *Errors for Height*. These cause inaccuracies in y_t^H and hence interfere with the ability of an O-SLDS to handle occlusions.

Scene A in particular suffers from all three problems due to the variety of furniture and their placement in the scene. In contrast, the partitions in Scene B produce particularly clean and abrupt occlusions, hence the O-SKF and O-SPF provide near-perfect results (see Figures 4g and 4h).

5.2 Single Camera View

In the case where only a single camera monitors the scene, an O-SLDS is still able to provide an improvement in tracking reliability. In a one-camera tracking situation, the number of false-positives will be minimal — every time the track is lost due to an occlusion, a new track will be formed

	Kalman Filter			Particle Filter			Obs-Sw. Kalman Filter			Obs-Sw. Particle Filter		
	TPs	*FPs*	*Precsn*	*TPs*	*FPs*	*Precsn*	*TPs*	*FPs*	*Precsn*	*TPs*	*FPs*	*Precsn*
Scene A	6242	4242	59.5%	6242	4274	59.4%	6242	1662	79.0%	6242	1412	81.6%
Scene B	2262	1857	54.9%	2297	1867	55.1%	2297	11	99.5%	2297	4	99.8%
Scene C	2657	3130	46.1%	2657	3093	46.2%	2657	498	84.2%	2657	544	82.8%

Table 1. Object count precisions of all evaluated filters. True-positive (TP) and false-positive (FP) counts are in frames.

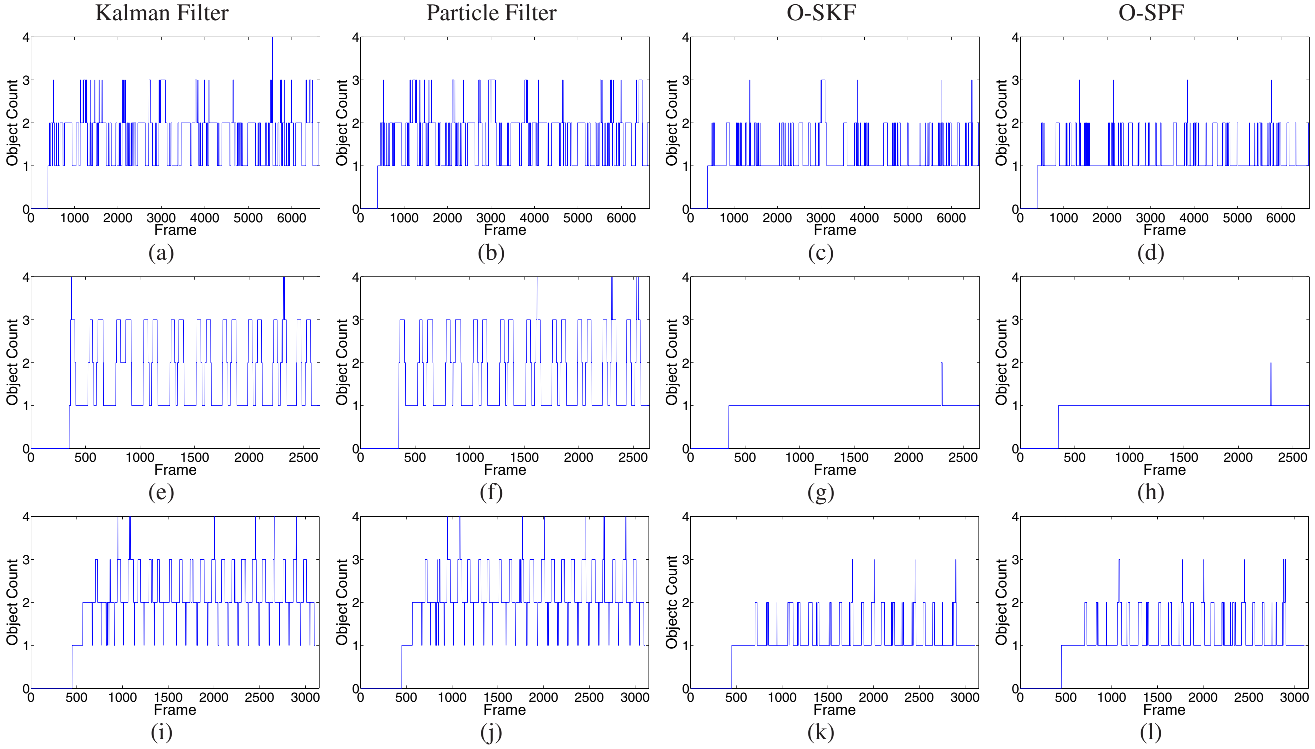
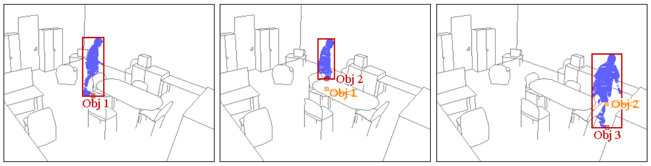


Figure 4. Tracked object counts for tracking in Scene A (top row), Scene B (middle row) and Scene C (bottom row) for the four filters tested. Note that the O-SKF and O-SPF significantly reduce false-positives (object counts over 1 are false-positives).

using the occluded blob. Hence evaluating track counts is inappropriate. Instead, this section considers the number of times that the person's track is lost and replaced by a new track. This will usually occur twice during a single occlusion — once at the beginning and once at the end (see Figure 5). However, an O-SLDS has the possibility of switching from tracking y_t^F to tracking y_t^H and thereby keeping a lock on the correct track throughout the occlusion.

Figure 6 plots the rate of lost tracks per occlusion. A rate of 1.0 indicates that one track was lost every time an occlusion begins or ends. The lost track rate thus indicates the proportion of occlusions that a filter fails to successfully track through. Note that the rate can exceed 1.0 due to multiple failures on a single occlusion or due to failures in background subtraction unrelated to occlusions.

As Figure 6 shows, the switching filters comprehensively outperform their standard counterparts when limited to single views. The two standard filters fail in three-quarters or more of occlusions, faring better in Scene A than Scenes B and C since occlusions by the partitions in Scenes B and C are always substantial and occur abruptly. In contrast, the O-SKF and O-SPF manage to successfully track through over 60% of all occlusions. Lost tracks still occur due to the fact that occlusions are not always 'clean' enough to ensure that the height is valid for correctly estimating y_t^H in a single monocular view. Furthermore, the lack of evidence from other views means that the correct track cannot be re-



Frame 440 Frame 450 Frame 496

Figure 5. Lost tracks typically occur in two places during each occlusion — once at the beginning and once at the end.

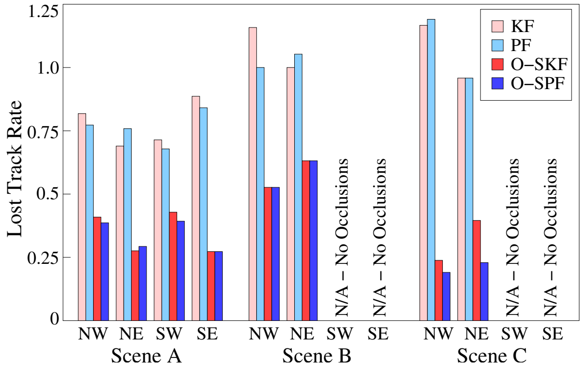


Figure 6. Rate of lost tracks per instance of occlusion for all filters when each view is processed independently.

covered after any initial error in tracking. For example, the O-SLDS filters perform most poorly in Scene B due to the fact that the partitions occlude the person each time they first enter the room. Without other views, there is no history of the person's true height and so there is no indication that an occlusion is occurring until the person moves past the occluding partition (whereupon the track is lost). When the person subsequently exits the room, their true height is known and the second occlusion is handled successfully.

Robustness could be increased by using higher-level reasoning to group together disjoint blobs from the same object, such as via appearance modelling [3]. Taken together, the group of disjoint blobs implies the true height of the person, alleviating many of the issues relating to blob splitting.

6 Conclusion

The concept of switching between two alternatives of the same measure (position) in an observation-switching linear dynamic system (O-SLDS) has been shown to improve the robustness of tracking in several realistic scenarios with frequent and significant occlusions by scene objects of the tracked person. Compared to unswitching filters, the approach can significantly decrease the occurrence of false-positive tracks in multi-camera environments and more than halve the number of lost tracks when using a single monocular camera. However, the system must observe the true height of the person *before* any occlusions occur and assumes that the person's head will be visible during an occlusion. Also, 'messy' occlusions that split up the person's blob can still confuse the system and cause tracking failures.

The approach is compatible with most linear dynamics models and is demonstrated with a switching Kalman filter and switching particle filter. Both types of filter benefit to a similar extent from the use of observation switching. Moreover, the improvements come without any prior knowledge of the scene or occluding obstacles — all occlusions are unexpected and an O-SLDS dynamically deals with the occlusions as they occur by switching between head- and foot-based measurements of the person's position.

Further robustness to occlusions could be achieved by combining the O-SLDS with complementary methods that predict occlusions based on scene information [3, 12]. Integration with approaches that handle tracking multiple mutually-occluding people would also be potential future work. In addition, the head and foot alternatives used in this paper could be extended to include more position measurements. These do not have to be limited to camera-based measurements. For example, range images can provide accurate location data but have slower frame rates than video and so could be interleaved with video-based measurements using the switching facility of the O-SLDS.

References

[1] Y. Bar-Shalom. *Tracking and Data Association*. Academic Press Inc., 1988.

[2] F. Fleuret, R. Lengagne, and P. Fua. Fixed point probability field for complex occlusion handling. In *IEEE Int'l Conf. on Computer Vision*, volume 1, pages 694–700, 2005.

[3] D. Greenhill, J. Renno, J. Orwell, and G. Jones. Occlusion analysis: Learning and utilising depth maps in object tracking. In *British Machine Vision Conf.*, 2004.

[4] W. E. L. Grimson, C. Stauffer, R. Romano, and L. Lee. Using adaptive tracking to classify and monitor activities in a site. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 22–29, 1998.

[5] M. Han, W. Xi, H. Tao, and Y. Gong. An algorithm for multiple object trajectory tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages I–865–871, 2004.

[6] K. P. Murphy. Switching Kalman filters. Technical report, University of California, Berkeley, August 1998.

[7] S. M. Oh, J. M. Rehg, T. Balch, and F. Dallaert. Learning and inference in parametric switching linear dynamic systems. In *IEEE Int'l Conf. on Computer Vision*, volume 2, pages 1161–1168, 2005.

[8] K. Otsuka and N. Mukawa. Multiview occlusion analysis for tracking densely populated objects based on 2-D visual angles. In *CVPR*, volume 1, pages 90–97, 2004.

[9] P. Peixoto, J. Batista, and H. Araújo. Real-time human activity monitoring exploring multiple vision sensors. In *Symposium on Intelligent Robotics Systems*, pages 221–228, 1999.

[10] R. Shumway and D. Stoffer. Dynamic linear models with switching. *Journal of the American Statistical Association*, 86:763–769, 1991.

[11] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[12] M. Xu and T. Ellis. Partial observation vs. blind tracking through occlusions. In *British Machine Vision Conf.*, 2002.

[13] T. Yang, S. Z. Li, Q. Pan, and J. Li. Real-time multiple objects tracking with occlusion handling in dynamic scenes. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages I–970–975, 2005.

[14] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1208–1221, September 2004.