Associating People Dropping off and Picking up Objects

Dima Damen David Hogg
School of Computing, University of Leeds
dima@comp.leeds.ac.uk, dch@comp.leeds.ac.uk

Abstract

Several interesting monitoring applications concern people entering a prescribed area, where they deposit an object in their possession, or collect an object deposited earlier. One example arises in the use of bicycle racks. We propose a novel method for associating each person who deposits an object with the person who later collects it. Our main contribution is to deal with ambiguity in the visual data through the use of global constraints on what is possible. The method is evaluated on a set of practical experiments in a bicycle rack, and applied to online theft detection by comparing the colour profile of associated individuals.

1 Introduction

A major challenge in computer vision is to reliably recognise events based on current object tracking and detection technology. Despite intensive research aiming towards universal tracking with "one object per track and one track per object" [17], such a tracker is not yet available if a single viewpoint is used. Ambiguous visual analysis makes it difficult to associate each person with the event accomplished. Understanding the expected events and their underlying global constraints is one way to resolve conflicting and ambiguous observations.

The scenario where a person leaves an object (typically locked) within a storage area, and picks it up sometime later, presents a rich constrained scenario that may be observed by a single CCTV camera. However, ambiguous observations from available tracking and object detection methods are often insufficient to recognize the events in isolation with any certainty. We propose a method to connect people and objects, and decide on the sequence of drop-off and pick-up events. The explanations generated should be both consistent and optimal based on the observations. A link can then be inferred between the person dropping off an object, and the person picking up the same object later. Additionally, passive biometric features can be utilized to compare these two individuals, and raise a warning when they do not match.

In Section 2, we give a brief overview of relevant work in associating trajectories and biometric comparison. Section 3 describes how we detect people and objects from a video stream. Section 4 presents the association task and proposes a solution method. Three experiments using bicycles as objects were conducted and results are presented in Section 5. We show how solutions are superior to those obtained when the constraints are relaxed or removed altogether.

BMVC 2007 doi:10.5244/C.21.9

2 Background

Our scenario requires associating the individual who drops off an object with the one who picks it up. Establishing correspondence between temporally disjoint motion trajectories of individuals has previously been used to relate entry and exit points in non-overlapping camera views, and to track individuals across blind regions with a single view [2, 3, 9, 12, 22]. Such methods typically depend on features of the moving individuals obtained from the video stream, sometimes referred to as passive [6] or soft-biometrics [19, 20].

Colour is a matching feature to connect two trajectories as it is easy to retrieve, although depends on people not changing their clothing within a single session [3, 6, 7, 9, 16, 21]. Bowden and KaewTraKulPong utilized colour histograms to re-identify individuals in non-overlapping neighbouring camera views [3]. ZhiHua and Komiya also used colour and shape for pedestrians and vehicles [22]. Similarly, Berclaz et. al. used both colour and location information to fuse trajectories in overlapping cameras [2]. Sivic et. al., used colour similarity of clothing to match people in family photos segmented using a face detector [16].

There are however difficulties in using colour information of walking pedestrians. Shadows, lighting changes and clothing's natural folding introduce wide variations to the colour of the corresponding pixels across frames. A pixel-by-pixel comparison can thus produce poor matches. Wu et. al. showed how frame-level and sequence-level colour representations can overcome pixel-level variations [20]. Frame-level colour information is usually represented by Gaussian mixtures [8] or colour histograms [3, 7, 14, 16, 21]. Bowden and KaewTraKulPong proposed the median histogram of the per-frame histograms to represent sequence-level information [3].

In our scenario, we can not ensure the picking person to be the same as the dropping person. Thus, correspondence based on the individual's features can not be applied. Makris et. al. followed an approach that is independent of feature-based matching. Their work learns the temporal characteristics of the source-sink connection by observing regularities of exits and entries in different camera views, and thereby anticipating the topology of a set of camera views [12]. Javed et. al. combine temporal relationships with colour profiles and neighbourhood knowledge. They concentrate on learning brightness transfer functions to strengthen their colour matching scheme [9]. The temporal expectancy can not be used in our case either. We can not anticipate with any certainty when an object will be picked up.

Kettnaker and Zabih, on the other hand, tracked pedestrians between cameras using a one-to-one optimal assignment approach [10]. This is based on the valid assumption that a trajectory can be associated with only one trajectory that left a neighbouring camera. We experiment with a similar technique to link dropping off and picking up trajectories.

To the authors' knowledge, the exact proposed scenario of dropping off and picking up objects has not been explored before. Perhaps the closest scenario that has been addressed previously is that of abandoned baggage. This was the basis of the PETS2006 challenge for which a number of solutions were proposed [5]. The abandoned baggage task, however, does not require associating events as required in dropping off and picking up objects.

The method described in this paper uses sequential propagation of multiple hypotheses in order to solve a constrained optimisation problem. A similar technique has been used in the radar surveillance literature to deal with ambiguities introduced by sensors.

Reid proposed the Multiple Hypothesis Tracking (MHT) in 1979 for radar activities [15], where tracking information, availed sequentially, clears ambiguities in previous observations. MHT keeps a set of hypotheses explaining legal assignments in a tree structure. Each child hypothesis represents one interpretation of the new data. A path in the tree represents a possible interpretation of all observations from the beginning of the observation period (root) up to the current timestamp (leaf).

3 Tracking People and Detecting Objects

For tracking people, a generic blob tracker [11] was used to retrieve the trajectories of individuals as they approach and depart the storage area. This tracker uses a per-pixel background model together with a simple foreground shape model, and assigns a unique identifier to each object moving over a continuous trajectory. For each person, the position (represented by centre of mass of the pixels), area (number of foreground pixels) and colour information are provided for each frame during the period the person remains visible. We extended the tracker to deal with broken trajectories through combining trajectories that exhibit similar colour profiles and that are spatially and temporally consistent.

In the case where the possessed object is comparable to the individual's projected area (such as bicycles utilized in our experiments), the change in area along the trajectory is significant, and can be used to differentiate people depositing from those collecting objects. This is feasible when the viewed locations are at similar depth from the camera's position. Figure 1 shows projected area through time as people deposit and collect objects. The difference in the area prior to entering the storage area and after leaving it represents an estimate of the projected size of an object at a given depth, and is independent of the person's size. Maximum likelihood estimation (MLE) was used to estimate Gaussian class conditional densities for area differences (at a given depth) across a drop-off and a pick-up (Figure 1). In this way we are able to estimate the likelihood that someone is dropping off or picking up an object, provided a continuous trajectory is available across the event. We use this likelihood, when available, to help constrain the assignment of individuals to different events detailed in Section 4.

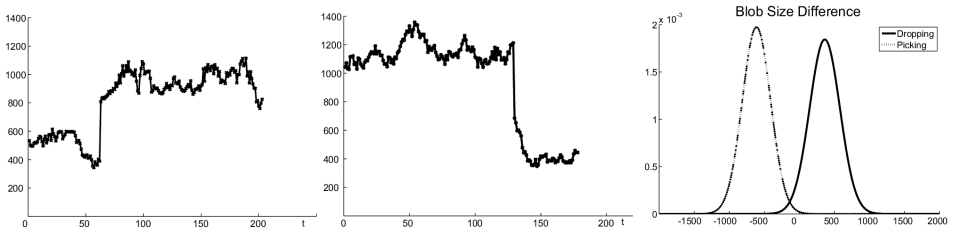


Figure 1: The change in area over time for someone picking (left) and dropping (middle) a bike; the increase/decrease in size is clearly visible around t=60 (left) and t=130 (middle), along with ML estimates for Gaussian distributions of area differences (right)

The tracker can not be used to identify static objects. Instead, 'before' and 'after' reference images of the storage area are compared, thereby revealing changed pixels, representing objects that have been deposited and removed. The risk of noise or lighting changes is minimised by taking reference images before a person enters the storage area

and after exiting it. If another person enters the area prior to the departure of the first, the second reference image is only taken after all have departed. We refer to these intervals as 'periods of activity'. The changed image pixels are then grouped into connected regions representing several blobs of dropped and picked objects as Figure 2 shows. We refer to these blobs as 'objects' even though some are actually multiple adjacent objects that could not be separated. The possibility of multiple objects of the same type within one detected blob is reflected in the constrained optimisation problem formulated in Section 4. We distinguish dropped from picked objects by detecting intensity edges in the 'before' and 'after' reference images, masking with the changed pixels, and classifying as a dropped object if the number of edge features increases, and vice versa. This assumes the background is relatively free of edge features, and that only one type of object (either dropped or picked) is present within each blob.

(a) (b) (c) (d)

Figure 2: Before (a) and after (b) reference images, revealing dropped (c) and picked (d) bicycles

4 The Association Phase

4.1 The problem as constrained optimisation

From tracking and object detection, a set of person hypotheses $\{p_i\}$ and a set of object hypotheses $\{o_i\}$ are generated. Individuals may appear more than once in the set of people, and, crucially, the same object is normally detected twice, once when it is dropped, and once when it is picked. False positives are expected in both sets. The association problem can be represented using a graph as in Figure 3. It connects people to objects and dropped objects to picked objects. There are two types of edges: person-object edges and object-object edges. The aim is to identify those edges that best explain the observations. An example of a solution is shown as darker lines to the right of Figure 3. Four types of events are used to explain the observations over an extended period:

1. $pkdp(p_i, o_j, p_k, o_l)$: person p_i picks up object o_j, person p_k drops off object o_l, o_j and o_l are the same object,
2. $dp(p_i, o_j)$: person p_i drops off object o_j, which remains unpicked during the period,
3. $pk(p_i, o_j)$: person p_i picks up object o_j, which was not dropped off during the period,
4. $none(p_i)$: person p_i neither picks up nor drops off an object.

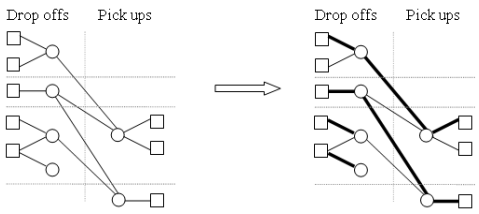


Figure 3: People (squares) and objects (circles) are connected via edges. Horizontal dividers separate periods of activity within the storage area, punctuated by periods of inactivity. The vertical divider separates drop-offs from pick-ups. Objects only appear as dropped or picked. People with broken trajectories can appear in both vertical sections

A possible explanation is specified by a union of events of each type:

$$e = C_{pkdp} \cup C_{dp} \cup C_{pk} \cup C_{none} \tag{1}$$

subject to the constraint that *each person is involved in exactly one event*. This is based on the assumption that each person does not drop/pick more than one object simultaneously. C_{pkdp} is the set of all (i, j, k, l) combinations that define the $pkdp$ events, and similarly for C_{dp}, C_{pk} and C_{none}.

An optimal solution e^* to the assignment problem is obtained by minimizing a cost function. This cost function is defined as a sum of the costs associated with each of the four event types:

$$\begin{aligned} f(e) &= \sum_{C_{pkdp}} f_{pkdp}(p_i, o_j, p_k, o_l) + \sum_{C_{dp}} f_{dp}(p_i, o_j) + \sum_{C_{pk}} f_{pk}(p_i, o_j) + \sum_{C_{none}} f_{none}(p_i) \\ f_{pkdp} &= d(p_i, o_j) + d(o_j, o_l) + d(p_k, o_l \mid o_j) \\ f_{dp} &= f_{pk} = d(p_i, o_i) + \alpha \\ f_{none} &= \beta \end{aligned} \tag{2}$$

where α is a penalty term on unconnected drop-off or pick-up events, and β is a penalty term on a 'none' event.

$d(p_i, o_j)$ is the plausibility of a link between a person and an object, and is derived from the maximum degree of overlap between the bounding box of the object and the bounding boxes of the person across the whole trajectory. If the trajectory of the person is complete, the dropping individuals (identified through area differencing - Section 3) are only connected to dropped objects, and the picking individuals are only connected to picked objects.

$$\begin{aligned} d(p_i, o_j) &= \begin{cases} 1 - \max\left(\frac{sharedBoundingBox(p_i, o_j)}{minBoundingBox(p_i, o_j)}\right) & \text{if } (I(p_i) \subset I(o_j)) \\ \infty & otherwise \end{cases} \\ I(p_i) &= [\text{time entering area}, \text{time exiting area}] \\ I(o_i) &= [\text{time of 'before' reference image}, \text{time of 'after' reference image}] \end{aligned} \tag{3}$$

$d(p_k, o_l \mid o_j)$ is an updated post-segmented cost. When two or more inseparable objects are added, one combined object blob is detected (Section 3). When one of these objects

is subsequently removed, a better estimate of the object's pixels and bounding box can be obtained as Figure 4 shows.

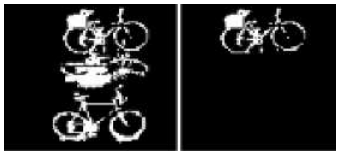


Figure 4: The left image shows three objects dropped simultaneously. As one object is collected (right image), post-segmentation could be achieved

$d(o_i, o_j)$ is the match of picked to dropped objects, and is assessed by comparing corresponding pixels. This match function accommodates any object type, and assumes objects do not change their shape or position between being dropped and picked. If $S(o_i)$ is the set of pixels representing object o_i, then

$$d(o_i, o_j) = \begin{cases} 1 - \frac{|S(o_i) \cap S(o_j)|}{\min(|S(o_i)|, |S(o_j)|)} & \text{if } o_i \in \text{picked} \wedge o_j \in \text{dropped} \wedge I(o_i) > I(o_j) \\ \infty & otherwise \end{cases} \quad (4)$$

4.2 Solving the constrained optimisation problem

To solve the constrained optimisation problem, we propagate a tree of multiple hypotheses (explanations) starting from the beginning of the observation period, and working through to the end, with levels of the tree corresponding to periods of activity. The tree is pruned at each stage to keep the search tractable (beam search) by retaining only the best hypotheses.

As several people enter the storage area simultaneously, several blobs get deposited or picked up. Multiple sets of assignments can be generated to explain the events during one period of activity, and decide who dropped/picked which object. If a person is connected to a picked object, then all matching drop offs that have not been picked up yet are compared for plausibility. The previously unmatched drop event is now joined with its pick up, and can not be picked again.

Each level in the tree is thus expanded into nodes representing the different hypotheses explaining the observations up to the current period of activity. Figure 5 shows a three-level multi-hypotheses tree. Each path (from root to leaf) in the tree corresponds to an explanation. The cost of the path equals the sum of the individual costs of events along that path (except dp events that are superseded by $pkdp$ events). The best path is determined by the minimum cost.

Due to the ambiguities in the visual data, the current best path may not be part of the best path to lower levels of the tree as it propagates into the future. Yet it would be impractical to maintain the complete tree, due to the number of possible hypotheses for all but the simplest cases. If l is the average number of sets of assignments per period of activity, and n is the number of such periods (levels of the tree), then the complexity will be $\Theta(l^n)$. For the first experiment which extended for one hour (refer to Section 5.1 for details), $l = 8.02$ and $n = 35$. This results in an intractable number of leaf hypotheses 4.43×10^{31} of which many are hypotheses of high cost. To ensure scalability over long video sequences involving many periods of activity, only the k-best hypotheses are

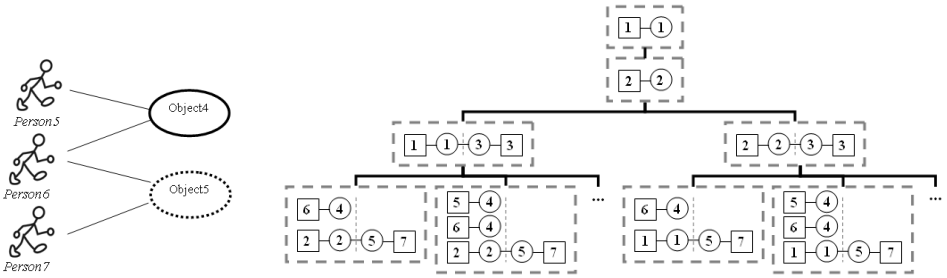


Figure 5: The left graph shows probable connections within a period of activity. Object4 represents a dropped object, while object5 represents a picked object. The various plausible explanations extend the tree. The event $pkdp(p_7, o_5, p_1, o_1)$ can not be added to the left branch as it contradicts the hypothesis at the parent level

retained at each level ($k = 10$ in our experiments). This reduces the complexity to $O(n)$. This method is 'on-line' in that the optimal hypothesis up to the current timestamp is always available.

For comparison purposes, we have implemented both an unconstrained solution (i.e. optimizing $f(e)$ free of the constraint) and one that only partially satisfies the constraint by transforming the problem into a one-to-one assignment problem between individuals and event types. This would still allow dropping people in the $pkdp$ events to be involved in another event. This approach is solved using the Munkres (Hungarian) algorithm. The cost matrix input to the assignment problem is an $n \times 4n$ matrix A where each person (row) can be mapped to another person or to one of the other three types of events explained above (columns). If one person can be matched to another person via more than one route, the minimum is selected, as Equation 5 shows.

$$A(i,k) = \min_{j,l} \{ f_{pkdp}(p_i, o_j, p_k, o_l) \} \tag{5}$$

5 Experiments and Results

5.1 Bicycles as Objects

The conducted experiments utilized bicycles as objects and bicycle racks as the storage area. One challenge in using bicycles is the overlap in their parking positions and the difficulty in distinguishing bicycles from each other. Three experiments were conducted. The first experiment extended over 1 hour and staged 28 different cyclists locking their bicycles onto the racks and picking them up again. A view from above was selected to avoid blocking the camera by passing pedestrians or vehicles (Figure 6). Cyclists were aware of the objectives of the research, but not of the techniques used in the solution. The second experiment was staged for 50 minutes and included 9 cyclists. The third full day (9.5 hours) experiment recorded 22 participants using the racks over the course of the day. In all three experiments, simulated thefts were performed. The number of staged thefts was 7, 1 and 5 respectively. Ground truth for the three datasets was established by hand, specifying the correct links between people and bicycles.

5.2 Association Results

The cost function (Equation 2) contains two parameters: α, β. The performance of the constrained solution is relatively insensitive to the value of these parameters, although the partially-constrained and unconstrained solutions are more sensitive to these values. We chose $\alpha = 1.5$ and $\beta = 2.0$ in the experiments presented here. Table 1 shows the percentage of people connected to the correct event in comparison to the ground truth. The constrained solution produced a significant improvement over the unconstrained or partially-constrained solutions. As expected, applying the constraint correctly connects a higher percentage of trajectories, due to the enforced uniqueness.

%	unconstrained	partially-constrained	constrained
exp1	75.86	86.21	93.10
exp2	70.37	70.37	92.59
exp3	83.59	82.03	96.09

Table 1: Percentage of correct connections

An example of an ambiguity that is resolved in the constrained solution from experiment 1 is shown in Figure 7 and Table 2. The unconstrained approach allowed p_{5387} to be linked as the dropping person twice, also p_{187} is involved in two events. The partially-constrained solution solved the problem for p_{5387}, but p_{187} is still involved in two events. The constrained solution satisfied the constraint fully producing the correct connections. Several similar cases across the three experiments could be found.

unconstrained	partially-constrained	constrained
none(187)	none(187)	
pkdp(2916, 3, 187, 1)	pk(2916, 3)	pkdp(2916, 3, 187, 1)
pkdp(8059, 15, 5387, 10)	pkdp(8059, 15, 5387, 10)	pkdp(8059, 15, 5387, 10)
pkdp(8215, 15, 5387, 10)	pkdp(8215, 15, 187, 1)	none(8215)

Table 2: Example from experiment 1 showing how adding constraints improves the associations. The notation for events is defined in Section 4

Figure 6: Viewpoint of the bicycle rack

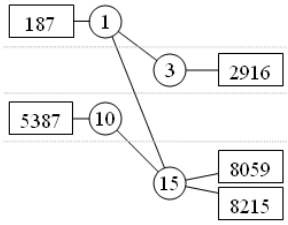
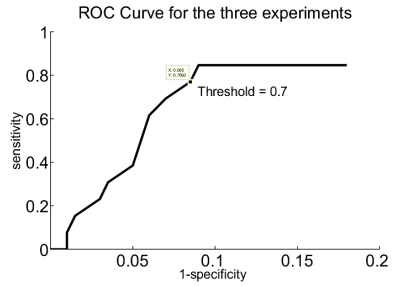


Figure 7: An example from experiment 1. Only the edges required for the example are included

5.3 Colour Comparison and Theft Detection

As an application of the method, we attempted to detect thefts of bicycles by comparing associated individuals using colour information. The presented results use the constrained solution. An $8 \times 8 \times 8$ scale-normalized equal-bin-size RGB colour histogram was generated from the foreground pixels at each frame. 'Scale-by-max' per channel was used as a simple colour constancy algorithm [1]. A per-bin median histogram was calculated across all frames as explained by Bowden and KaewTraKulPong [3]. A distance metric between histograms was produced using histogram intersection [18].



	Predicted	
Actual	Thief	Non-Thief
Thief	10	3
Non-Thief	17	183

Figure 8: ROC curve (left) representing theft detection results. 0.7 was selected as the threshold to calculate the confusion matrix (right) for the three experiments

The ROC curve is shown in Figure 8. At a threshold of 0.7, 77% (10 out of 13) of the theft cases were caught for an 8.5% false-positive rate. 4 false-positive cases resulted from the owner returning wearing different clothing, demonstrating the limitation of using only colour profiles. 4 other false-positive cases were incorrectly connected, while 9 were correctly connected but the colour comparison failed to match the individuals due to poor segmentation from the background.

6 Conclusions and Future Work

The paper has proposed a method for associating individuals as they drop objects off and pick them up sometime later using an online constraint-based optimisation. Ambiguities in the observations are expressed as multiple hypotheses, which can then be verified or invalidated by future observations. Experiments proved the value of the association framework for bicycle theft detection, where colour profiles were used to compare linked individuals. To strengthen the system, the person's features beyond colour could be added to compare individuals, for example, height [4], body mass, gait [13] and behaviour analysis.

Though our experiments were confined to bicycles and bicycle racks, the approach could be applied in other contexts. Car parks, cloakrooms and other parking environments exhibit analogous events and constraints and might also be viable. Although the techniques used to track and detect objects vary across domains, the propagation of multiple hypotheses used to solve the constrained optimisation should remain applicable.

References

[1] K. Barnard, V. Cardei, and B. Funt. A comparison of computational color constancy algorithms—part 1: Methodology and experiments with synthesized data. *IEEE Trans. Image*

Processing, 11(9):972–983, 2002.

[2] J. Berclaz, F. Fleuret, and P. Fua. Robust people tracking with global trajectory optimization. In *Proc. CVPR*, volume 1, pages 744–750, 2006.

[3] R. Bowden and P. KaewTraKulPong. Towards automated wide area visual surveillance: tracking objects between spatially-separated, uncalibrated views. *Proc. Vision, Image and Signal Processing*, 152(2):213–223, 2005.

[4] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. In *Proc. ICCV*, volume 1, pages 434–441, 1999.

[5] J. Ferryman, editor. *Ninth IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance (PETS2006)*. IEEE, New York, 2006.

[6] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *Proc. CVPR*, volume 2, pages 1528–1535, 2006.

[7] A. Gilbert and R. Bowden. Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In *Proc. ECCV*, pages 125–136, 2006.

[8] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank. Principal axis-based correspondence between multiple cameras for people tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(4):663–671, 2006.

[9] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *Proc. CVPR*, volume 2, pages 26–33, 2005.

[10] V. Kettnaker and R. Zabih. Bayesian multi-camera surveillance. In *Proc. CVPR*, volume 2, pages 259 – 265, 1999.

[11] D. Magee. Tracking multiple vehicles using foreground, background and motion models. In *Proc. ECCV Workshop on Statistical Methods in Video Processing*, pages 7–12, 2002.

[12] D. Makris, T. Ellis, and J. Black. Bridging the gaps between cameras. In *Proc. CVPR*, volume 2, pages 205–210, 2004.

[13] M. Nixon and J. Carter. Automatic recognition by gait. *Proceedings of the IEEE*, 94(11):2013–2024, 2006.

[14] M. Piccardi and E. D. Cheng. Track matching over disjoint camera views based on an incremental major color spectrum histogram. In *IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 147–152, 2005.

[15] D. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24(6):843–854, 1979.

[16] J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *Proc. BMVC*, volume 3, pages 909–918, 2006.

[17] C. Stauffer. Learning to track objects through unobserved regions. In *IEEE Workshop on Motion and Video Computing*, volume 2, pages 96–102, 2005.

[18] M. Swain and D. Ballard. Color indexing. *Int. Journal of Computer Vision*, 7(1):11–32, 1991.

[19] Y. Wang, E. Chang, and K. Cheng. A video analysis framework for soft biometry security surveillance. In *Proc. ACM Int. workshop on Video surveillance & sensor networks*, 2005.

[20] G. Wu, A. Rahimi, E. Y. Chang, G. Kingshy, T. Tsai, J. Ankur, and Y. Wang. Identifying color in motion in video sensors. In *Proc. CVPR*, pages 561–569, 2006.

[21] W. Zajdel, Z. Zivkovic, and B. J. A. Krose. Keeping track of humans: Have i seen this person before? In *Proc. ICRA*, pages 2081–2086, 2005.

[22] L. ZhiHua and K. Komiya. Region-wide automatic visual search and pursuit surveillance system of vehicles and people using networked intelligent cameras. In *Proc. Int. Conference on Signal Processing Proceedings*, volume 2, pages 945–8, 2002.