

# Multi-View Video Analysis of Humans and Vehicles in an Unconstrained Environment

D.M. Hansen<sup>†</sup>, P.T. Duizer<sup>†</sup>, S. Park<sup>‡</sup>, T.B. Moeslund<sup>†</sup>, and M.M. Trivedi<sup>‡</sup>

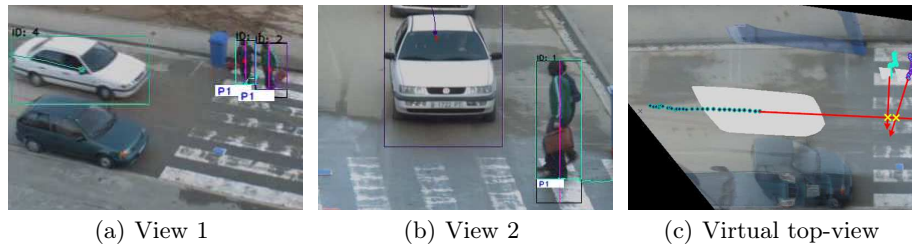
<sup>†</sup>Computer Vision and Media Technology, Aalborg University, Denmark

<sup>‡</sup>Computer Vision and Robotics Research, University of California, San Diego

**Abstract.** This paper presents an automatic visual analysis system for simultaneously tracking humans and vehicles using multiple cameras in an unconstrained outdoor environment. The system establishes correspondence between views using a principal axis approach for humans and a footage region approach for vehicles. Novel methods for locating humans in groups and solving ambiguity when matching vehicles across views are presented. Foreground segmentation for each view is performed using the codebook method and HSV shadow suppression. The tracking of objects is performed in each view, and occlusion situations are resolved by probabilistic appearance models. The system is tested on hours of video and on three different datasets.

## 1 Introduction

Visual analysis of motion has received much interest in recent years for its practical importance in various application areas [1, 2]. The majority of work on visual analysis has focused on either monitoring humans or vehicles, but the amount of work focusing on monitoring humans *and* vehicles simultaneously is relatively small. The simultaneous tracking of humans and vehicles is often desirable for enhanced situational awareness in real-world environments. The motions of humans and vehicles may involve different characteristics, which may request different analysis techniques. This paper presents an automatic visual surveillance system for simultaneously tracking humans and vehicles in unconstrained outdoor environments by using multiple cameras. Our system uses overlapping-view cameras to exploit the merit of multi-view approach to extract view-invariant features of objects (i.e., persons and vehicles) such as view-invariant size, position, and velocity of the objects in the real world environment. Given such view-invariant information, it would be possible to develop more enhanced ambient intelligent systems and intelligent infrastructure. For example, visual surveillance systems could capture the movements of vehicles and pedestrians and predict possible collisions, as illustrated in Figure 1. Furthermore, the motion-patterns of vehicles and pedestrians could be analyzed to detect abnormal behavior, e.g., drunk driving. Such incidents could control the traffic lights or directly communicate with the control of the vehicle. A prerequisite for such systems is robust detection and tracking of humans and vehicles, which is the focus of this paper.



**Fig. 1.** Collision detection in a multi-view sequence. The trajectories are mapped to a virtual top-view for visualization purposes. A red arrow depicts the extended velocity vectors of the tracked objects, and the yellow crosses show intersections as an indication of a possible collision. The gray area in the right-most image shows the positions or footage region of the objects, calculated as the area on the ground plane covered in both views by the object. Note that only parts of the images are shown in order to increase visibility.

## 2 Previous work

Many multi-view systems have been proposed and they can be categorized into two groups: disjoint camera views [3, 4] *vs.* overlapping camera views [5, 6]. While the disjoint views are effective for covering wide field of views, the overlapping views are desirable for efficient handling of severe occlusions which inevitably occur in unconstrained urban environments. The overlapping views furthermore improve the accuracy in estimating the position and size of objects.

The object matching between multiple views can be achieved by: recognition based methods [7, 8] or geometry based methods [6, 9], or a combination of these [10]. When the views are not overlapping, the color histograms can be compared between different views for object matching [4], while the geometry-based methods [11] are preferred when the views are overlapping.

Some approaches have used full camera calibration for accuracy [6, 9], whereas other approaches in more recent works adopted homography mapping for versatility [12, 13]. The full camera calibration based methods fused observations from the multiple views to produce fairly precise matching; however, the calibration is a resource-consuming process and a skill-demanding task in practice [2], especially in outdoor scenes. The homography-based methods [12, 13] establish the matching between the multiple views by using the so-called ‘principal axis correspondence’ and epipolar geometry. Tracking objects can be enhanced by the particle filter or the Kalman filter [13] for tracking. The above systems aimed mainly at human tracking in multiple views. Some other homography-based method uses a graph-cut algorithm to segment and track each object [14].

One of the few systems that do multi-view tracking of both humans and vehicles is [15], where objects are tracked using their footage region. Objects occupying a larger area on the ground plane results in a larger footage region in the homography domain. However, since humans are significantly smaller than vehicles the footage region is not robust for tracking humans. The works in [14,

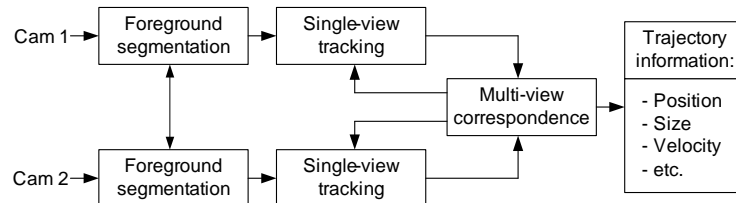
10] present good results using the footage region for tracking humans, but only in a constrained environment.

Other systems that track both humans and vehicles are [16] and [17]. [16] calculates each object's geolocation for camera handoff, but the system requires a three-dimensional site model which is difficult to obtain in practice. In contrast [17] does not require camera calibration, however, this system does not use overlapping cameras and does not produce an output suitable for interaction analysis.

In conclusion, the systems summarized above do not provide an efficient solution for tracking of both humans and vehicles. Our proposed system combines the principal axis and footage region approaches in an integrated manner for establishing a better foundation in simultaneously tracking and analyzing the interactions between humans and vehicles in real-world busy traffic scenes.

### 3 System Overview

The system developed in this work consists of three modules, as depicted in Figure 2. For each synchronized camera input moving objects are segmented using the *foreground segmentation* module (Section 4). The moving objects are tracked using the *single-view tracking* module (Section 5). Following this, the tracks are matched across views using the *multi-view correspondence* (Section 6). Feedback from this module is used to improve the accuracy of the single-view tracking. The output from the system is a trajectory of each detected and tracked object. Since the system is quite comprehensive we have primarily focused on describing the novel contribution of this work, i.e., multi-view correspondence and evaluation on very long and realistic sequences.



**Fig. 2.** System overview.

### 4 Foreground Segmentation

The foreground segmentation allows the forthcoming modules to focus their attention on areas containing objects. An unconstrained environment present many challenges such as, background camouflage, reflection, illumination change,

shadow, dynamic background and changing background. The foreground segmentation is based on motion segmentation followed by shadow suppression. Motion segmentation is based on a robust background subtraction method which dynamically update the background model and add new background layers. We have in earlier work [18] showed the robustness of this approach in very long and unconstrained videos sequences.

A problem left unsolved by the motion segmentation is moving cast shadows. Results of cast shadows are false shape, size and appearance of objects and merge of otherwise separate foreground objects. For shadow suppression an HSV color segmentation approach based on [19] is used. However, since this method is based solely on color segmentation, it is not optimal for separating cast shadow from self shadow, which is a part of the object. We therefore enhance the method with multi-view information to reduce the number of falsely classified shadow pixels. Concretely we consider the footage region, see section 6.2, as potential shadow. This improves the segmentation since the footage region in many cases only contains the cast shadow and not the object. For details see [20].

## 5 Single-view Tracking

The single-view tracking serves two purposes: 1) to classify objects as either human or vehicle and 2) to track each single object through the scene.

We classify an object as either a *human* or a *vehicle* based on how steep the vertical projection of the foreground mask is using a spread measure from [12]. Though this measure is not view invariant, a threshold of 0.08-0.10 is sufficient to separate humans from vehicles. Furthermore, as in [12], humans walking in a group are split using two thresholds to find peaks and valleys in the vertical projection histogram. The detected object is split at each valley point.

The tracking of objects are, whenever possible, done using the bounding boxes together with a Kalman filtering approach and some smoothing constraints, see [20] for details. To handle more complicated cases like occlusion we use an approach based on probabilistic appearance models [21].

Each track has its own probabilistic appearance model, which consists of an RGB color model with an associated probability mask. The use of probabilistic appearance models can be viewed as weighted template matching, where the template is an appearance model and the weights are given by the associated probability mask. The coordinates of the model are normalized to the object centroid.

For each new track, a new probabilistic appearance model is created. A track refinement step is applied before updating the model at each match by finding the best fit in a small neighborhood e.g.  $5 \times 5$  pixels. Track refinement increases the accuracy of the model. When updating, the model usually stabilizes after half a second at 15 fps. Detail on building and applying the model can be found in [21].

## 6 Multi-view Correspondence

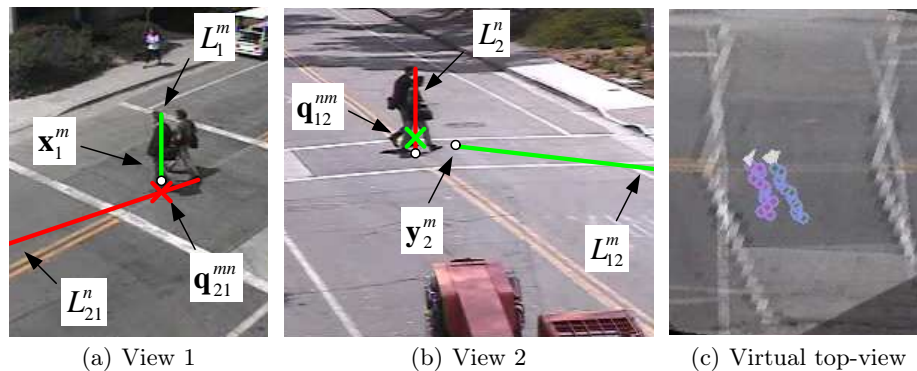
The correspondence module match objects tracked within the overlapping region of the cameras. Multi-view correspondence is needed to handle difficult occlusion cases, e.g. during full occlusion or at initialization where the probabilistic appearance models are not reliable. Since humans and vehicles are very different object types, two separate procedures are followed.

### 6.1 Correspondence of Humans

A principal axis is a vertical line in the image domain, which is fitted to each tracked human using least median of squares. Assuming a homography between views, the principal axes can be mapped from one view to the other. Analyzing the intersections between the principal axes in one view and the principal axes mapped into this view allows for finding correspondences between views [12].

With this approach two or more human objects are often tracked as one object, especially when people enter the scene as a group, even though they are separable in one view. Other work like [21] and [12] “wait” until a group split before tracking individuals. However, based on observations, people that enter as a group are most likely to stay as a group while moving through the scene.

To enable correspondence of people/individuals in groups, a novel correspondence algorithm is developed in this work, which is executed after the correspondence algorithm from [12]. The new algorithm locates unpaired tracks in view 1, and test if they match an already paired track in view 2. If the matching satisfies two distance constraints the unpaired track in view 1 is paired with the paired track in view 2. The steps of the algorithm are explained in the following, with Figure 3 as illustration.



**Fig. 3.** Resolving the problem of a group being tracked as a single object in view 2. See text for explanation. Note that only parts of the images are shown in order to increase visibility.

1. A list ( $\theta_1^{unpaired}$ ) of all unpaired principal axes in view 1 is created and a list ( $\theta_2^{paired}$ ) of all paired principal axes in view 2 is created.
2. A principal axis ( $L_1^m$ ) in  $\theta_1^{unpaired}$  is selected and compared with all principal axes in  $\theta_2^{paired}$ , in this case ( $L_2^n$ ). For each comparison two distances are calculated:
 

**Distance 1:**  $L_2^n$  is mapped from view 2 into view 1 ( $L_{21}^n$ ) and the intersection point is found ( $\mathbf{q}_{21}^{mn}$ ). The distance from the intersection point to person  $m$ 's ground point location is found as  $D_1 = |\mathbf{q}_{21}^{mn} - \mathbf{x}_1^m|$ . The distance is calculated this way for comparison with distance 2; the two distances are summed in step 3.

**Distance 2:**  $L_1^m$  is mapped from view 1 into view 2 ( $L_{12}^m$ ) and the intersection point is found ( $\mathbf{q}_{12}^{nm}$ ). Furthermore, the ground point location,  $\mathbf{x}_1^m$ , is also mapped from view 1 into view 2,  $\mathbf{y}_2^m$ . The distance from the intersection point to the mapped ground point is found as  $D_2 = |\mathbf{q}_{12}^{nm} - \mathbf{y}_2^m|$ .
3. If  $D_1 < D_{T1}$  and  $D_2 < D_{T2}$ , then  $n$  along with the score  $D_1 + D_2$  are stored in the list  $A_m$ .  $D_{T1}$  and  $D_{T2}$  are selected empirically.
4. Select person  $n$  with the smallest score in  $A_m$  and label the pair  $(m, n)$ .
5. Remove  $L_1^m$  from  $\theta_1^{unpaired}$ . Go to step 2 and repeat the procedure with a different  $m$  until  $\theta_1^{unpaired}$  is empty.

The algorithm is executed a second time with the order of view 1 and view 2 switched.

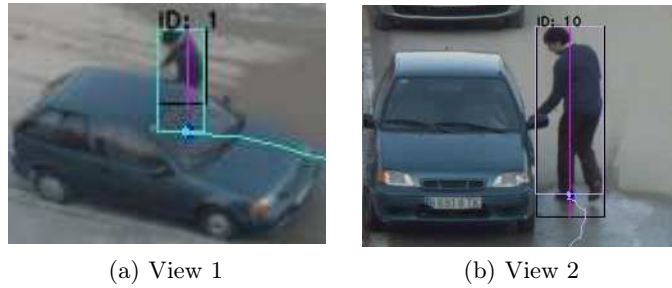
Compared to [12] this is a greedy algorithm and not a global algorithm. The distance  $D_2$  is applied, because it is expected that the mapped ground point should be close to the intersection of the principal axes in view 2.

Using the new correspondence algorithm it is possible to find the ground point location of a human within a group, even when it is tracked as a single object. This is fed back to the single-view tracking module where it is used to correct the track, especially the lower part of the bounding box and hence the centroid. In Figure 4 an example is provided.

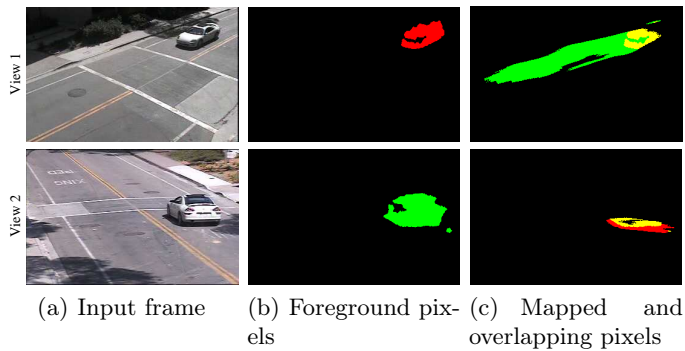
## 6.2 Correspondence of Vehicles

The correspondence of vehicles utilizes the footage region, as in work by [15] and [14]. However, [15] and [14] does not apply single view tracking; both methods map all foreground pixels into a common domain and perform tracking in this domain. By tracking in each view we are able to maintain an object's track even when it is missing in one of the views, and still maintain the benefit of improved accuracy when the object is detected in both views.

Our approach has two main steps. The first is to create an overlap matrix and, secondly, to solve any ambiguity expressed in the overlap matrix. The overlap matrix is created by mapping all foreground pixels belonging to a vehicle track from view 1 into view 2 and find overlap with any foreground pixels for a vehicle track in view 2. The principle is illustrated in Figure 5, where a vehicle is visible in both views. In the overlap matrix three scenarios are possible: One-to-one,



**Fig. 4.** Tracking of occluded person. Despite being occluded by the vehicle, the person's ground point is correctly located. View 1 is zoomed in on the person. A purple vertical line is a principal axis. The black box and the colored box indicate the bounding box location before and after performing correspondence.



**Fig. 5.** Finding overlap between vehicle tracks in view 1 and view 2. If there is overlapping pixels (marked by yellow), the tracks are a possible pair.

Many-to-one, and Many-to-many. *One-to-one* is a straight forward situation with no ambiguity, see Figure 5. *Many-to-many* overlap occurs when two vehicles drive by each other and at some point occlude each other in one view. As a result the foreground mask of one of the vehicles will wrongly overlap with both vehicles in the other view. *Many-to-one* overlap (or one-to-many) could be caused by two vehicles being tracked as one in a view, or the view with many vehicle tracks could wrongly track noise classified as a vehicle. The many-to-one and many-to-many overlaps are considered as ambiguity. When solving it, the vehicle track with the most possible overlaps is solved first, etc. In the end, only one-to-one overlaps are left.

One of two methods is applied to solve this ambiguity in both the many-to-many and many-to-one situation. The first method ensures that a historic relationship is preserved between vehicle tracks. If the vehicles have been corresponded correctly before the occlusion, the ambiguity is solved by maintaining the historic correspondence.

However, it is not guaranteed that the historic relationship is available, e.g. at track initialization. In these situations a different approach based on a “plausible ground point” is applied. Taking the vehicle’s centroid and a vertical line through the centroid, the plausible ground point is located at the lowest foreground pixels on this vertical line, as illustrated in Figure 6. The plausible ground point is located on the ground, but also beneath the vehicle itself. When mapping this plausible ground point into the other view using the planar homography, the mapped point should therefore also be located beneath the car in this view. The mapped point is used to solve the ambiguity and is illustrated in Figure 6. The plausible ground point is only applied if the history is not reliable. After

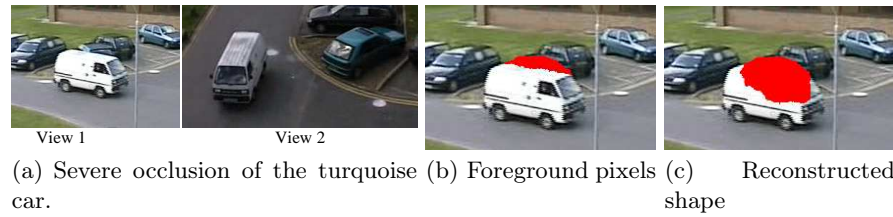


**Fig. 6.** The plausible ground point can be used to solve ambiguities in the overlap matrix. The many-to-many overlap occurs in this situation because the vehicle with red centroid in view 1 occludes the lower part of the vehicle with the green centroid.

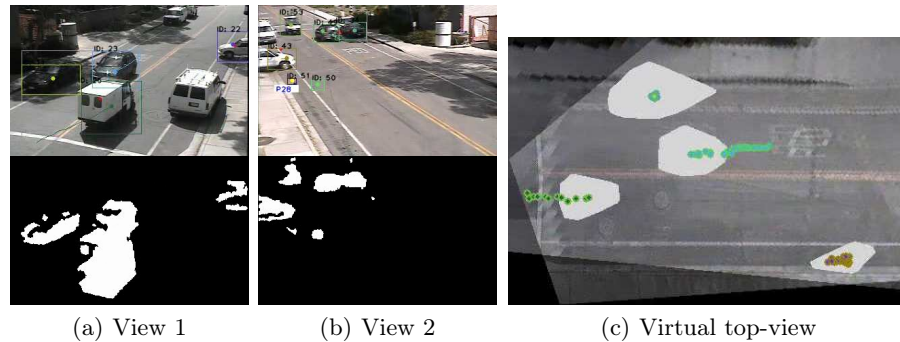
resolving the ambiguities and pairing the vehicle tracks, it might happen that a pair of vehicle tracks does not overlap even though it is historically expected. The missing pair of vehicle tracks could be caused by the entire bottom portion of the vehicle being occluded in one view. An example of this is shown in Figure 7(a), where the white van is occluding the bottom portion of the turquoise car in view 1. Before occlusion, the turquoise car has been paired correctly between views and a pairing is therefore expected. The foreground pixels assigned to the turquoise car in view 1 are shown in Figure 7(b) during occlusion. To solve this problem, the foreground mask is reconstructed from the probabilistic appearance models which holds a memory of the shape of the turquoise car, as shown in Figure 7(c). With the reconstructed shape it is again possible to find overlap. The view invariant representations for four vehicles are shown in Figure 8.

## 7 System Evaluation

The primary test of this system is performed using our own dataset, since this is much longer than public datasets. The cameras have a resolution of  $352 \times 240$  pixels. Inter-object occlusion and illumination changes frequent occur in the dataset. To ensure realistic results the system was running continuously with



**Fig. 7.** Solving severe occlusion by reconstructing the shape of the occluded vehicle. Note that only parts of the images are shown in order to increase visibility.

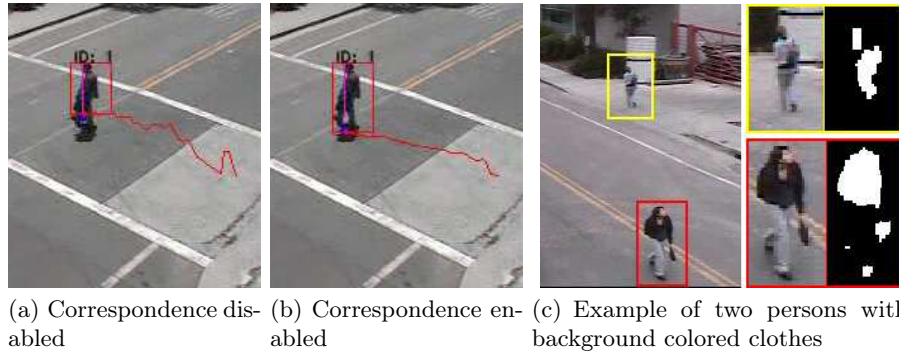


**Fig. 8.** Correct tracking of four moving vehicles, the white van is not detected since it is parked and is part of the background model. Vehicles are represented by their footage region in the virtual top-view. To avoid “holes” in the footage region caused by imperfect foreground segmentation a convex hull is fitted to the region. The point location of each vehicle is given by the centroid of the convex hull area. The noise detected in view 2 is not tracked in the virtual top-view since there is no corresponding object in view 1. Note that (c) has been cropped in order to increase visibility.

the same parameters for more than 67 hours. During this time we selected six sequences covering different times of the day (at different days). In total, 385 minutes of test data. This contains 1351 humans and 267 vehicles<sup>1</sup>.

If a track is missing in both or one view, or is wrongly corresponded or classified for more than 0.5 seconds it is considered an error. With this definition the percentage of correctly tracked objects is 67.9% with 2.3% non-existing objects being tracked. However, in most cases where a part of a track is missing it is maintained correctly in one of the views. This makes this type of error less serious and by ignoring it the percentage of correctly tracked objects increases to 88.2%. In contrast to single view tracking systems, this system handles occlusion very well, see Figure 8 and 10, and locates objects with high accuracy. In Figure 9 it is illustrated how the accuracy is improved using correspondence between views. The most significant cause of error in this test is the foreground

<sup>1</sup> Note that the system runs in near real-time on a 3Ghz computer with 2GB ram. When many large objects are present the framerate drops to around 10Hz.



**Fig. 9.** (a) and (b): Example of the increased accuracy gained from doing correspondence for humans. (c): The RGB values of the foreground objects are very similar to the background; only the intensity sets them apart which makes them very hard to segment correctly. Note that only parts of the images are shown in order to increase visibility.

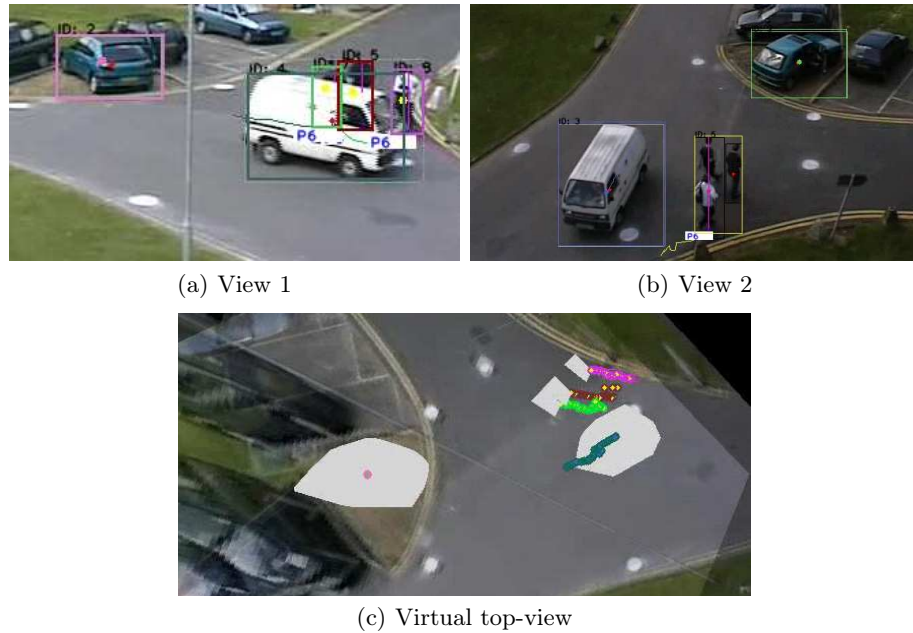
segmentation. Because the system is tested with long sequences, the quality of the foreground segmentation is not optimal, which cause a drop in performance. This leads to wrong object classification and track initialization especially for vehicles. The main problem with humans is that they are removed as noise when the misdetection is too severe.

The most frequent cause of error in the foreground segmentation is background camouflage as exemplified in Figure 9(c). Furthermore, the shadow suppression parameters are hard to adjust for segmentation over day long periods.

### 7.1 Test on PETS'2001 dataset

The system is tested on Dataset 1 and 2 of the PETS'2001 dataset. In Dataset 1, the only error that occurs is a track switch in a group with three people as they walk behind a recently parked car. All remaining humans and vehicles are tracked correctly. The only error that occurs in Dataset 2 is a person that is not detected properly due to his resemblance with the cars and road in the background. Otherwise the tracking is correct.

An example from the PETS'2001 dataset is show in Figure 10. The three persons in view 2 are automatically initialized and tracked, see Figure 10(c), using the group correspondence algorithm presented in Section 6.1. This would not be possible without the group correspondence since the three persons are tracked as one as they enter, and the vertical projection can not be used to separate them. Other systems that track humans using the footage region like [15] and [14] would not be able to handle this situation, since the inter-object occlusion in both views would result in misshaped and wrongly connected footage regions. In fact this is in general the case for systems testing on this particular part of the PETS'2001 dataset, unless tracking is manually initialized.



**Fig. 10.** Example of tracking in the PETS'2001 dataset. Three persons enter the scene as a group. The group correspondence algorithm initialize and track all three persons correctly even during occlusion by the white van. Note that groups are only given one ID, but the system still holds information about all three individuals as seen by the trajectories in the virtual top-view. Note also that (c) has been cropped in order to increase visibility.

## 8 Conclusion

We have presented a robust system for tracking humans and vehicles through their activities and interactions in an unconstrained outdoor environment using multiple surveillance cameras. To our knowledge, very little work has been done on traffic monitoring of both humans and vehicles, since most related work focuses on either humans *or* vehicles.

The proposed system integrates novel versions of the principal axis-based approach and the footage region-based approach for simultaneously establishing the multi-view correspondence of humans and vehicles. The system robustly handles severe occlusions, efficiently locates individual humans in groups, and handles ambiguity in correspondences between vehicles.

We conducted extensive experimental evaluations with long-term videos captured from unconstrained outdoor environments in different sites in different days. The experimental evaluation shows the efficacy of the system, and the proposed system can be used as a tracking module for higher-level behavior analysis of persons and vehicles such as the estimation of collision likelihood between humans and vehicles.

## References

1. T.B. Moeslund, A. Hilton and V. Kruger, "A Survey of Advances in Vision-Based Human Motion Capture and Analysis", *CVIU*, Vol. 104(2-3), pp. 90-126, 2006.
2. M. Valera and S.A. Velastin, "Intelligent Distributed Surveillance Systems: A Review", *Vision, Image, and Signal Processing*, Vol. 152(2), pp. 192-204, 2005.
3. O. Javed, Z. Rasheed, K. Shafique and M. Shah, "Tracking in Multiple Cameras with Disjoint Views", *ICCV* 2003.
4. F. Porikli and A. Divakaran, "Multi-Camera Calibration, Object Tracking and Query Generation", *Int. Conference on Multimedia and Expo*, Washington, DC, 2003.
5. C. Stauffer and K. Tieu, "Automated Multi-Camera Planar Tracking Correspondence Modeling", *CVPR* 2003.
6. A. Mittal and L. S. Davis, "M2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene", *IJCV*, Vol. 51(3), pp. 189-203, 2003.
7. J. Orwell, P. Remagnino and G. A. Jones, "Multi-Camera Colour Tracking", *IEEE Int. Workshop on Visual Surveillance*, Vol. 28(4), Fort Collins, CO, USA, 1999.
8. A. Gilbert and R. Bowden, "Incremental, Scalable tracking of Objects Inter Camera", *CVIU*, 111(1), pp. 43-58, 2008.
9. R. Pflugfelder and H. Bischof, "People Tracking Across Two Distant Self-Calibrated Cameras", *IEEE Conference on Advanced Video and Signal based Surveillance*, London, UK, 2007.
10. F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multi-Camera People Tracking with a Probabilistic Occupancy Map", *PAMI*, 30(2), pp. 267-282, 2008.
11. S. Calderara, R. Cucchiara, and A. Prati, "Bayesian-Competitive Consistent Labeling for People Surveillance", *PAMI*, 30(2), pp. 354-360, 2008.
12. W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou and S. Maybank, "Principal Axis-Based Correspondence between Multiple Cameras for People Tracking", *PAMI*, Vol. 28(4), pp. 625-634, 2006.
13. S. Park and M. M. Trivedi, "Understanding Human Interactions with Track and Body Synergies (TBS) Captured from Multiple Views", *CVIU*, 111(1), pp. 2-20, 2008.
14. S. M. Khan and M. Shah, "A Multiview Approach to Tracking People in Crowded Scenes using a Planar Homography Constraint", *ECCV* 2006.
15. S. Park and M. M. Trivedi, "Analysis and Query of Person-Vehicle Interactions in Homography Domain", *IEEE Conference on Video Surveillance and Sensor Networks*, Santa Barbara, CA, USA, 2006.
16. R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt and L. Wixson, "A System for Video Surveillance and Monitoring", *Tech. report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University*, May, 2000.
17. M. Shah, O. Javed and K. Shafique, "Automated Visual Surveillance in Realistic Scenarios", *IEEE MultiMedia*, Vol. 14(1), pp. 30-39, 2007.
18. P. Fihl, R. Corlin, S. Park, T.B. Moeslund, and M.M. Trivedi, "Tracking of Individuals in Very Long Video Sequences", *International Symposium on Visual Computing*, Lake Tahoe, Nevada, USA, LNCS 4291, 2006.
19. R. Cucchiara, C. Grana, M. Piccardi, A. Prati and S. Sirotti, "Improving Shadow Suppression in Moving Object Detection with HSV Color Information", *IEEE Conference on Intelligent Transportation Systems*, Oakland, CA, USA, 2001.
20. D.M. Hansen and P.T. Duizer, "Multi-View Video Surveillance of Outdoor Traffic", *Master Thesis*, Aalborg University, Denmark, 2007.
21. A. Senior, A. Hampapur, Y.-L. Tian, L. Brown, S. Pankanti and R. Bolle, "Appearance Models for Occlusion Handling", *IVC*, Vol. 24(11), pp. 1233-1243, 2006.