Computer Vision Winter Workshop 2006, Ondřej Chum, Vojtěch Franc (eds.)
Telč, Czech Republic, February 6–8
Czech Pattern Recognition Society

Multi-Camera Tracking for Visual Surveillance Applications

David Thirde[1], Mark Borg[1], James Ferryman[1], Josep Aguilera[2], Martin Kampel[2], and Gustavo Fernandez[3]

[1] Computational Vision Group, The University of Reading, UK
{D.J.Thirde, M.Borg, J.Ferryman}@rdg.ac.uk
[2] Pattern Recognition and Image Processing Group, Vienna University of Technology, Austria
{agu,kampel}@prip.tuwien.ac.at
[3] Video & Safety Systems, ARC Seibersdorf Research GmbH, Austria
Gustavo.Fernandez@arcs.ac.at

Abstract *This paper presents the multi-camera tracking component of a complete surveillance system that was developed as part of the AVITRACK project, involving 10 academic and industrial partners. The aim of the project is to automatically recognise activities around a parked aircraft in an airport apron area to improve the efficiency, safety and security of the servicing operation. The multi-camera tracking module takes as input per-camera tracking and recognition results and fuses these into object estimates using a common spatio-temporal co-ordinate frame. The multi-camera localisation and tracking of objects is evaluated for a range of test data.*

1 Introduction

This paper details tracking work undertaken on the EU project AVITRACK. The aim of the AVITRACK project is to automatically recognise activities around a parked aircraft in an airport apron area to improve the efficiency, safety and security of the operation. A combination of visual surveillance and event recognition algorithms are applied in a decentralised multi-camera end-to-end system providing real-time recognition of the activities and interactions of numerous vehicles and personnel in a dynamic environment. A main requirement behind the adopted architecture is that the implemented system must be capable of monitoring and recognising the apron activities over extended periods of time, operating in real-time (12.5 FPS, colour, PAL resolution). In this paper we discuss the multiple-camera object tracking system, the output of this system is a set of estimated objects on the airport apron.

The tracking system discussed in this paper comprises per-camera (2D) video frame tracking and multi-camera (3D) fused object tracking. Video frame tracking methods generally require methods to detect and track the objects of interest. Motion detection methods attempt to locate connected regions of pixels that represent the moving objects within the scene; there are many ways to achieve this including frame to frame differencing, background subtraction and motion analysis (e.g. optical flow) techniques. Image plane based object tracking methods take as input the result from the motion detection stage and commonly apply trajectory or appearance analysis to predict, associate and update previously observed objects in the current time step. The tracking algorithms have to deal with motion detection errors and complex object interactions in the congested apron area e.g. merging, occlusion, fragmentation, non-rigid motion, etc. Apron analysis presents further challenges due to the size of the tracked objects with prolonged occlusions occuring frequently throughout apron operations. To recognise the tracked objects in the observed scene both top-down(e.g. [8]) and bottom-up methods(e.g. [6]) can be applied. The challenges faced in apron monitoring are the quantity (28 categories) and similarity of objects to be classified e.g. many vehicles have similar appearance and size.

Multi-camera tracking combines the data measured by the individual cameras to maximise the useful information content of the observed apron. The main challenge for apron monitoring is the tracking of large objects with significant size, existing methods generally assume point sources [3] and therefore extra descriptors are required to improve the association. People entering and exiting vehicles also pose a problem in that the objects are only partially visible therefore they cannot be localised using the ground plane.

The paper is organised as follows: Section 2 gives an overview of the Scene Tracking module, Section 3 gives an overview of the per-camera tracking, Section 4 describes the multi-camera tracking and Section 5 gives experimental results showing the performance of the localisation and tracking performance.

2 Scene Tracking

The AVITRACK Scene Tracking module is responsible for the estimation of objects on the airport apron in two distinct stages — per camera (2D) object tracking and centralised world (3D) object tracking. The per camera object tracking (Section 3) consists of motion detection to find the moving objects in the observed scene, followed by tracking in the image plane of the camera. The tracked objects are subsequently classified using a hierarchical object recognition scheme. The tracking results from the eight cameras are then sent to a central server where the multiple observations are fused into single estimates (Section 4).

3 Per-Camera Tracking

The per-camera tracking sub-module is used to extract the moving objects of interest using spatio-temporal video processing. The first task of the per-camera tracking sub-module is to perform motion detection to segment a video image into connected regions of foreground pixels that represent moving objects. These results are then used to track objects of interest across multiple frames. For AVITRACK, a total of 16 motion detection algorithms were implemented and quantitatively evaluated on various apron sequences under different environmental conditions. After taking into account the evaluation results [2], the colour mean and variance method [9] was used for AVITRACK.

Real-time object tracking can be described as a correspondence problem, and involves finding which object in a video frame relates to which object in the next frame. Normally, the time interval between two successive frames is small, therefore inter-frame changes are limited, thus allowing the use of temporal constraints and object features to simplify the correspondence problem. The Kanade-Lucas-Tomasi (KLT) feature tracking algorithm [7] is used for tracking objects – this combines a local feature selection criterion with feature-based matching in adjacent frames. As the KLT algorithm considers features to be independent entities and tracks each of them individually, to be able to move from the feature-tracking level to the object-tracking level, the KLT algorithm is incorporated into a higher-level tracking process that groups features into objects, maintain associations between them, and uses the individual feature-tracking results to track objects through complex object interactions. Object interaction during this matching process is particularly challenging for the apron environment because of the diversity in object sizes and because most of the activity occurs near the aircraft with extended periods of merging and occlusion.

To improve reasoning in the multi-camera tracking module, we introduce a confidence measure that the 2-D measurement represents the whole object. Localisation is generally inaccurate when clipping occurs at the left, bottom or right-hand image borders when objects enter/exit the scene. The confidence measure $\psi \in [0, 1]$ is estimated using a linear ramp function at the image borders (with $\psi = 1$ representing 'confident' i.e. the object is unlikely to be clipped). A single confidence estimate ψ_{O_i} for an object O_i is computed as a product over the processed bounding box edges for each object.

To efficiently recognise the people and vehicles on the apron, a hierarchical approach is applied that comprises both bottom-up and top-down classification. The first stage categorises the top-level types of object that are expected to be found on the apron (people, ground vehicle, aircraft or equipment); this is achieved using a bottom-up Gaussian mixture model classifier trained on efficient descriptors such as 3D width, 3D height, dispersedness and aspect ratio. This was inspired by the work of Collins *et al* [6] where it was shown to work well for distinct object classes. After the first coarse classification, the second stage of the classification is applied to the vehicle category to recognise the individual sub-types of vehicle. Such sub-types cannot be determined from simple descriptors and hence a proven method is used [8] to fit textured 3D models to the detected objects in the scene. Because model-based classification is computationally intensive, the algorithm runs on a background thread and is synchronised with the bottom-up classifier through a processing queue. Some latency occurs in object localisation, but this is corrected later on in the multi-camera tracking module.

4 Multi-Camera Tracking

The multi-camera object tracking component of the overall tracking system takes as input the tracked objects from each of the eight cameras and outputs estimates of the object location, velocity, orientation and size in the 3D world co-ordinate system. Spatial registration of the cameras is performed using per camera coplanar calibration and the camera streams are synchronised temporally across the network by a central video server.

4.1 Object Localisation

The localisation of an object in the context of visual surveillance generally relates to finding a location in the world co-ordinates that is most representative of that object. This is commonly taken to be the centre of gravity of the object on the ground plane and it is this definition that we adopt here. With accurate classification and detection, the localisation of vehicles in the 3D world can be reduced to a 2D geometrical problem. For state of the art algorithms accurate classification and detection is not reliable enough to apply such principled methods with confidence. For the AVITRACK project we therefore devised a simple, but effective, vehicle localisation strategy that gives good performance over a wide range of conditions.

The first step of the strategy is to categorise the detected objects as *person* or *non-person* using a supervised Gaussian mixture model of the estimated object width and height in world co-ordinates. The motivation behind this is that people generally have negligible depth compared to vehicles and hence a different strategy is required to locate each type. For the person class of objects the location is taken to be the bottom-centre of the bounding box of the detected object, this location estimate for people is commonplace in visual surveillance systems.

For vehicles many researchers arbitrarily choose the centroid of the bounding box / detected pixels to locate the object in the world. This method has the drawback that for objects further away from the camera the bottom of the bounding box is a better approximation of the object location than the centroid. To alleviate this problem we compute the angle made between the camera and the object to estimate an improved location. For a camera lying on the ground plane the location of the object will be reasonably proximal to the bottom centre of the bounding box, whereas for an object viewed directly overhead the location of the object will be closer to the measured centre of the bounding box.

Using this observation we formulated a smooth function to estimate the position of the centroid using the (2-D) angle to the object. Taking α to be the angle mea-

sured between the camera and the object, the proportion p of the vertical bounding box height (where $0 \leq p \leq 1/2$) was estimated as $p = 1/2(1 - \exp(-\lambda a))$; the parameter $\lambda \equiv ln(2)/(0.15 \times 1/2\pi)$ was determined experimentally to provide good performance over a range of test data. The vertical estimate of the object location was therfore taken to be $y_lo + p * h$ where y_lo is the bottom edge of the bounding box and h is the height of the bounding box. The horizontal estimate of the object location was measured as the horizontal centre-line of the bounding box, since this is generally a reasonable estimate.

4.2 Data Association Filter

Three distinct methods have been evaluated to perform data association between existing tracks and observations. Two of these are ground-plane based i.e. they constrain all objects to be located on the ground-plane. The association methods used in conjunction with the ground-plane constraint are the nearest neighbour and joint probabilistic association filters [3], both these are implemented to use a Kalman filter to estimate the state of the observed objects. The third association method uses the epipolar constraint to detect objects that are off the ground plane (or partially visible). To improve the data association in congested apron regions the validation mechanism has been extended to incorporate information about the object trajectory and categorisation information. Finally, the associated data is either fused into a single estimated observation or used to update the Kalman filter in a sequential manner.

Nearest Neighbour Data Association (NNDA) Filter The data association step associates existing track predictions with the per camera measurements. In the nearest neighbour filter the nearest match within a validation gate is determined to be the sole observation for a given camera. For multiple tracks viewed from multiple sensors the nearest neighbour filter is:

1. Obtain the validated set of measurements for each track.
2. Associate the nearest measurement with each track, repeated for all cameras.
3. Fuse associated measurements over all cameras into a single measurement.
4. Kalman filter update of each track state with the fused measurement.

Joint Probabilistic Data Association (JPDA) Filter The discrete nature of the NNDA filter leads to a degradation of performance in the prescence of noise, where the chance of mis-association is increased. To improve the robustness in the presence of noise the JPDA filter analyses the neighbourhood of the track estimate to compute the joint probability of association between each measurement and track combination. Briefly, the JPDA filter is performed as follows (see [3] for further details):

1. Cluster tracks into extended validation regions using the intersection of validation gates.
2. For each extended validation region, generate all feasible hypotheses of track to measurement associations. The feasibility constraint requires that each track generates at most one measurement and that each measurement corresponds to only one track.
3. Compute the probabilities of the feasible hypotheses
4. Find the association probability between a track and a measurement by summating the hypothesis probabilities for all hypotheses in which the measurement occurs.
5. Compute the combined innovation for use in the sequential Kalman filter update using the standard PDA filter expressions.

Epipolar Data Association (EDA) Filter To track objects that cannot be located on the ground plane we have extended the tracker to perform epipolar data association (based on the method presented in [4]), this can either be run in standalone mode or as an extension to the ground-plane tracking system. The epipolar data association method is a technique for associating per-camera observations *independent* of the existing objects. This method is performed as follows:

1. Associate per-camera observations using the epipolar plane constraint.
2. The associated measurements are formed into fused observations using a method based on the covariance intersection approach, the estimated intersection point of the epipolar lines is used to locate the fused observation.
3. Associate the fused observations with the existing tracks (since this relationship is not known), this is achieved using a variant of the NNDA filter.

Validation of Measurements The validated set of measurements are extracted using a validation gate [3], this is applied to limit the potential matches between existing tracks and observations. In previous tracking work the gate generally represents the uncertainty in the spatial location of the object; in apron analysis this strategy often fails when large and small objects are interacting in close proximity on the congested apron, the uncertainty of the measurement is greater for larger objects hence using spatial proximity alone larger objects can often be mis-associated with the small tracks. To circumvent this problem we have extended the validation gate to incorporate velocity and category information, allowing greater discrimination when associating tracks and observations.

The observed measurement is a 7-D vector:

$$\mathbf{Z} = [x, y, \dot{x}, \dot{y}, P(p), P(v), P(a)]^T$$

where $P(\cdot)$ is the probability estimate that the object is one of three main taxonomic categories (p = Person, v = Vehicle, a = Aircraft). This extended gate allows objects to be validated based on spatial location, motion and category, which improves the accuracy in congested apron regions. The effective volume of the gate is determined by a threshold τ

on the normalised innovation squared distance between the predicted track states and the observed measurements:

$$d_k^2(i,j) = \left[\mathbf{H}\widehat{\mathbf{X}}_k^-(i) - \mathbf{Z}_k(j)\right]^T \mathbf{S}_k^{-1} \left[\mathbf{H}\widehat{\mathbf{X}}_k^-(i) - \mathbf{Z}_k(j)\right] \quad (1)$$

where $\mathbf{S}_k = \mathbf{H}\widehat{\mathbf{P}}_k^-(i)\mathbf{H}^T + \mathbf{R}_k(j)$ is the innovation covariance between the track and the measurement; this takes the form:

$$\mathbf{S}_k = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} & 0 & 0 & 0 & 0 & 0 \\ \sigma_{yx} & \sigma_y^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{\dot{x}}^2 & \sigma_{\dot{x}\dot{y}} & 0 & 0 & 0 \\ 0 & 0 & \sigma_{\dot{y}\dot{x}} & \sigma_{\dot{y}}^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{P(p)}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_{P(v)}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_{P(a)}^2 \end{bmatrix} \quad (2)$$

For the kinematic terms the predicted state uncertainty $\widehat{\mathbf{P}}_k^-$ is taken from the Kalman filter and constant *a priori* estimates are used for the probability terms. Similarly, the measurement noise covariance $\mathbf{R}$ is estimated for the kinematic terms by propagating a nominal image plane uncertainty into the world co-ordinate system using the method presented in [4]. Measurement noise for the probability terms is determined *a priori*. An appropriate gate threshold can be determined from tables of the chi-square distribution [3]. For epipolar data association the validation gate also includes estimated height information.

Data Fusion, Track Maintenance and Contextual Information In the NNDA and EDA filters the matched observations are combined to find the fused estimate of the object, this is achieved using covariance intersection. This method estimates the fused uncertainty $\mathbf{R}_{fused}$ for N matched observations as a weighted summation:

$$\mathbf{R}_{fused} = \left(w_1\mathbf{R}_1^{-1} + \ldots + w_N\mathbf{R}_{numcams}^{-1}\right)^{-1} \quad (3)$$

where $w_i = w_i'/\sum_{j=1}^{N} w_j'$ and $w_i' = \psi_i^c$ is the confidence of the i'th associated observation (made by camera c) estimated using the method in Section 3. Sequential Kalman filter update was used in the JPDA filter to estimate the object states from associated measurements.

If tracks are not associated using the extended validation gate the requirements are relaxed such that objects with inaccurate velocity or category measurements can still be associated. Remaining unassociated measurements are fused into new tracks, using a validation gate between observations to constrain the association and fusion steps. Ghosts tracks without supporting observations are terminated after a predetermined period of time. For certain events predefined contextual rules have to be added to the tracking module e.g. during aircraft arrival a global association of the per-camera aircraft tracks is made to circumvent the problem that no single camera observes the whole object.

5 Experimental Results

The evaluation of the tracking system assesses the performance of the multi-camera tracking on representative test data. This evaluation focusses on object localisation and the multi-camera tracking algorithms presented in Section 4. An overview of the evaluation of the per-camera tracking and recognition modules is given in [1]

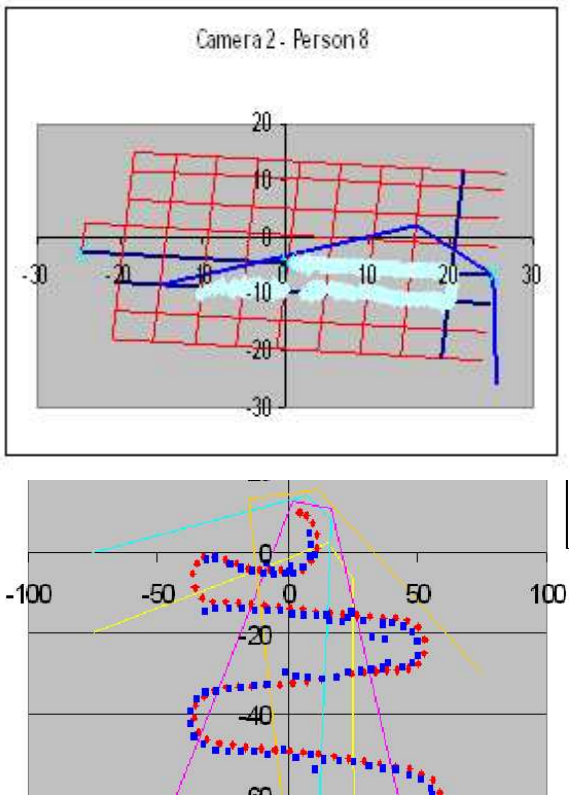

Figure 1: (Top) 2D trajectory graph for a person (S27, camera 2). (Bottom) 2D trajectory graph for a vehicle (S27, cameras 3,4,5,6) showing (Red) the EGNOS trajectory and (Blue) the estimated location on the apron (the scale is in metres). Both examples show the camera fi elds of view.

5.1 Object Localisation

For the evaluation of the 3D localisation module individual person and vehicle tracks have been considered. The estimated path of the objects is compared to a preset path. For each location along the path the shortest Euclidean distance (in metres) is computed between the point and the associated grid line. More detail on the evaluation of object localisation is given in [5].

For the person class, it can be seen that person trajectory in Figure 1 is broken due to occlusions. Occlusions lead to loss of 3D data information causing errors on 3D trajectory reconstruction. It was found that the accuracy of the person localisation is approximately 1 metre average over all cameras, this is to be expected due to detection or calibration error. Due to the general innaccuracy in the far-field of all cameras these results show that the use of multiple overlapping cameras is justified for this surveillance system to ensure the objects are accurately located on the apron.

The evaluation of the vehicle localisation was performed on a dataset for which EGNOS positional measurements

were recorded[1] The evaluation is performed using a single trajectory estimate made by the camera with the largest viewable object. The reasoning for this is that the EGNOS data was captured over a large area, and several cameras can view this trajectory. The results, also shown in Figure 1, demonstrate that the estimated vehicle location is reasonably accurate close to the camera sensors (at the top of the figure). In the far field the estimate diverges from the measured EGNOS signal due to the perspective effect and the uniform quantisation of the sensor pixels. Quantatively, the mean distance between the EGNOS signal and the estimated location was found to be 2.65 metres $+/-0.34$. The minimum deviation was found to be 0.58 metres and the maximum was found to be 4.64 metres.

5.2 Multi-Camera Tracking with Extended Validation Gates

The extension of the NNDA filter with an improved validation criteria and measurement confidence is qualitatively evaluated for two representative test sequences: S21 (9100 frames) containing people and vehicles moving over the apron area and S28 (1200 frames) containing a crowded scene with many objects interacting within close proximity near the aircraft.

The performance is shown in Figure 2 where estimated objects on the ground plane are shown for the two test sequences. It is clear to see that by extending the validation gate to include velocity and category, as well as the use of measurement confidence in the fusion process, the extended NNDA filter out-performs the standard (i.e. spatial validation and fusion) process. Many more objects estimated by the extended filter are contiguous, with less fragmentation and more robust matching between measurements and existing tracks. It can be seen that the extended filter is robust against objects that are not on the ground-plane (e.g. the containers on the loader in S28). This is achieved by using camera line-of-sight to determine that the container observations do not agree between the cameras and hence the estimated object is given a lower confidence.

The results are encouraging, for many scenarios the extension of the validation gate provides much greater stability, especially when objects are interacting in close proximity. It is noted that the track identity can be lost when the object motion is not well modelled by the Kalman filter or when tracks are associated with spurious measurements. The filter currently has no contextual information about the 3D geometry of the scene, therefore the camera line-of-sight cannot be accurately determined. Due to this factor, objects can have lower than expected confidence since some camera measurements cannot be made due to occlusions. The addition of contextual information would also allow the tracking of large objects when they are off the ground-plane (e.g. the containers in S28). For larger objects epipolar analysis is not practical, therefore contextual information about the loader vehicle would be required to position the container objects correctly.

[1]The EGNOS measurements were kindly provided by the ESA project GAMMA (http://www.m3systems.net/project/gamma/); the EGNOS system gives an estimated accuracy of 2-3m for 95% of measurements.

6 Conclusions and Future Work

This paper discusses and extends multi-camera tracking algorithms with applicability to real-time surveillance systems. The tracking performance has been evaluated in the context of airport activity monitoring, demonstrating both the effectiveness and robustness of the algorithms. The multi-camera tracking system will be further extended to include a particle filter based approach, once this intergration is complete a comparative evalution will be performed against ground-truth data. Further to this the full tracking system will be assessed at the airport to validate its robustness during live 24/7 operation.

References

[1] J. Aguilera, D. Thirde, M. Kampel, M. Borg, G. Fernandez, and J. Ferryman. Visual surveillance for airport monitoring applications. In *11th Computer Vision Winter Workshop 2006*, Czech Republic, 2006.

[2] J. Aguilera, H. Wildernauer, M. Kampel, M. Borg, D. Thirde, and J. Ferryman. Evaluation of motion segmentation quality for aircraft activity surveillances. In *Proc. Joint IEEE Int. Workshop on VS-PETS, Beijing*, Oct 2005.

[3] Y. Bar-Shalom and X. Li. *Multitarget-Multisensor Tracking: Principles and Techniques*. YBS Publishing, 1995.

[4] J. Black and T. Ellis. Multi Camera Image Measurement and Correspondence. In *Measurement - Journal of the International Measurement Confederation*, volume 35 num 1, pages 61–71, 2002.

[5] M. Borg, D. Thirde, J. Ferryman, K. Baker, J. Aguilera, and M. Kampel. Evaluation of object tracking for aircraft activity surveillance. In *Proc. Joint IEEE Int. Workshop on VS-PETS, Beijing*, Oct 2005.

[6] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, and L. Wixson. A System for Videosurveillance and Monitoring: VSAM Final Report. Technical Report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University, May 2000.

[7] J. Shi and C. Tomasi. Good features to track. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[8] G. Sullivan. Visual interpretation of known objects in constrained scenes. In *Phil. Trans. R. Soc. Lon.*, volume B, 337, pages 361–370, 1992.

[9] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfi nder: Real-time tracking of the human body. In *IEEE Transactions on PAMI*, volume 19 num 7, pages 780–785, 1997.

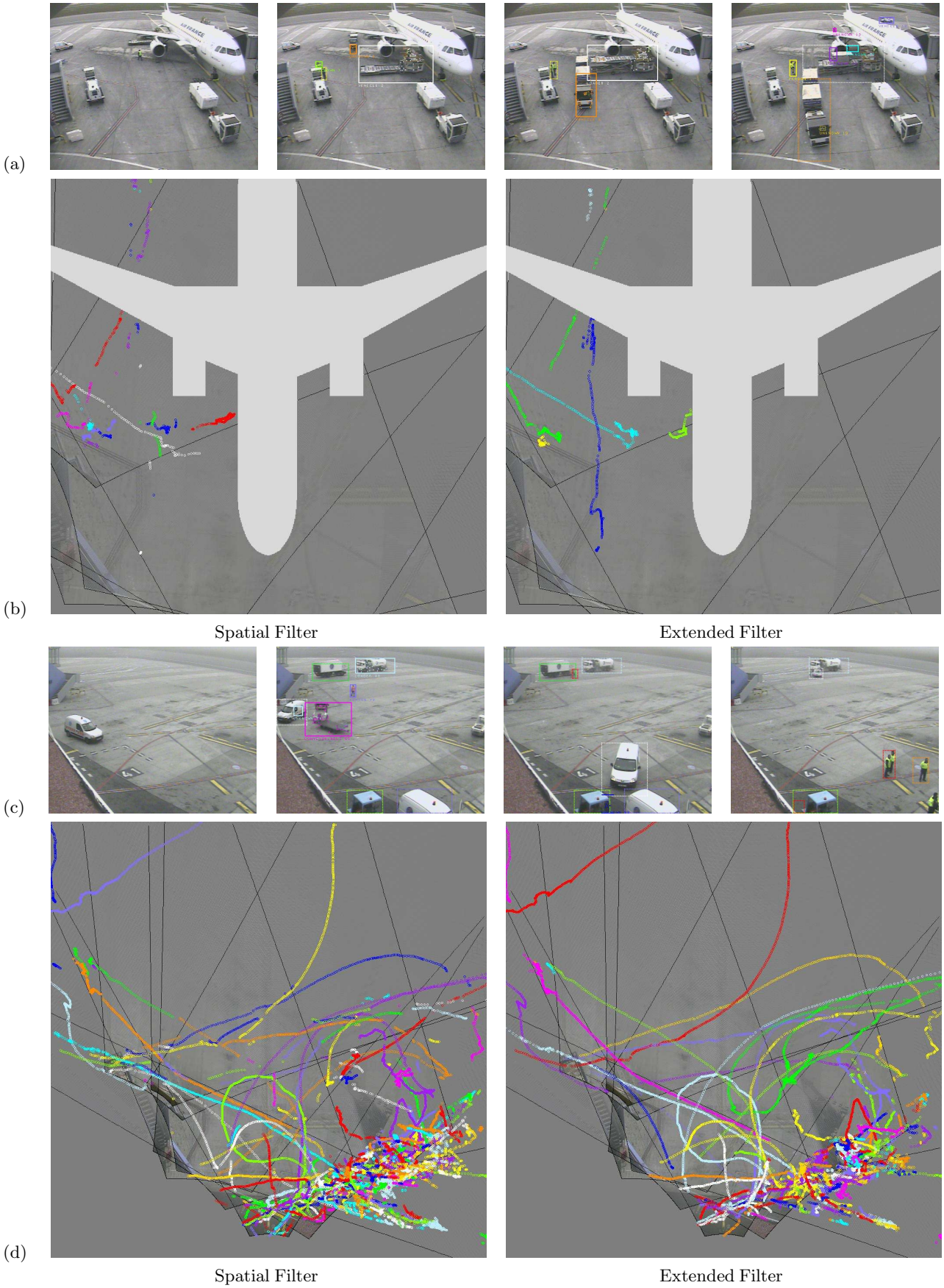

Figure 2: Results of the multi-camera tracking module showing extracted object locations on the ground-plane for two data sets. The track colour is derived from the object ID, limited to eight colours for visualisation. (a) S28 - All cameras frames 0, 500, 750, 1000. (b) Objects tracked by the NNDA filter with (Extended Filter) and without (Spatial Filter) the extended validation gate and confidence based fusion. The aircraft is added for illustrative purposes. (c) S21 - All cameras frames 0, 6000, 7000, 9000. (d) Objects tracked by the NNDA filter with (Extended Filter) and without (Spatial Filter) the extended validation gate and confidence based fusion.