Distributed Multi-Camera Surveillance for Aircraft Servicing Operations

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

[1] Computational Vision Group, The University of Reading, UK
{D.J.Thirde, M.Borg, J.M.Ferryman}@reading.ac.uk
http://www.cvg.reading.ac.uk/
[2] Pattern Recognition and Image Processing Group
Vienna University of Technology, Austria
{agu,kampel}@prip.tuwien.ac.at
http://prip.tuwien.ac.at/

Abstract

This paper presents the visual surveillance aspects of a distributed intelligent system that has been developed in the context of aircraft activity monitoring. The overall tracking system comprises three main modules — Motion Detection, Object Tracking and Data Fusion. In this paper we primarily focus on the object tracking and data fusion modules.

1 Introduction

This paper describes work undertaken on the EU project AVITRACK[3]. The main aim of this project is to automate the supervision of commercial aircraft servicing operations on the ground at airports (in bounded areas known as *aprons*). A combination of visual surveillance algorithms are applied in a decentralised multi-camera environment with overlapping fields of view (FOV) [1] to track objects and recognise activities predefined by a set of servicing operations. Each camera agent performs per frame detection and tracking of scene objects, and the output data is transmitted to a central server where data association and fused object tracking is performed. The system must be capable of monitoring a dynamic environment over an extended period of time, and must operate in real-time (defined as 12.5 FPS with resolution 720×576) on colour video streams.

The tracking of moving objects on the apron has previously been performed using a top-down model based approach [2] although such methods are generally computationally expensive. An alternative approach, bottom-up scene tracking, refers to a process that comprises the two sub-processes *motion detection* and *object tracking*; the advantage of bottom-up scene tracking is that it is more generic and computationally efficient compared to the top-down method.

[3] This work is supported by the EU, grant AVITRACK (AST3-CT-3002-502818).

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. Background subtraction methods, such as [3], store an estimate of the static scene, which can be accumulated over a period of observation; this background model is subsequently applied to find foreground (i.e. moving) regions that do not match the static scene.

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 vehicles tracked, therefore prolonged occlusions occur frequently throughout apron operations. The Kanade-Lucas-Tomasi (KLT) feature tracker [4] combines a local feature selection criterion with feature-based matching in adjacent frames; this method has the advantage that objects can be tracked through partial occlusion when only a sub-set of the features are visible. To improve the computational efficiency of the tracker motion segmentation is not performed globally to detect the objects. Instead, the features are used in conjunction with a rule based approach to correspond connected foreground regions; in this way the KLT tracker *simultaneously* solves the problems of data association and tracking without presumption of a global motion for each object.

The data fusion module combines tracking data seen by each of the individual cameras to maximise the useful information content of the observed apron. The main challenge of data fusion for apron monitoring is the tracking of large objects with significant size, existing methods generally assume point sources [1] 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.

In this paper, Section 2 introduces the use of visual surveillance in ambient intelligence systems. Section 3 reviews the per camera motion detection, objects tracking and categorisation. Section 4 describes the data fusion module and Section 5 contains evaluation of the presented methods.

2 Visual Surveillance for Ambient Intelligence

A real-time cognitive ambient intelligence (AmI) system requires the capability to interpret pervasive data arising from real-world events and processes acquired from distributed multimodal sensors. The processing systems local to each sensor require the capability to improve the estimation of the real-world events by sharing information. Finally, this information is shared with the end users, suggesting decisions and communicating through human terms to support them in their tasks. The work presented on the AVITRACK project represents the initial steps in the development of such a system, with intelligent interpretation of the scene via distributed vision based agents. In the longer term it is anticipated that the vision based agents will be able to share information with e.g. GNSS location agents, PTZ camera agents, infra-red camera agents, radar-based agents and RFID tag agents etc. The sharing of information between multimodal sensors provides a more accurate, more complete, representation of the events as they unfold in the scene.

The long term aim of airport surveillance in this context is to provide the end users with the capability to use the information distributed by the AmI system. The cognition of human actions through aural, visual and neural sensors coupled with intelligent processing is a fundamental part of such a system since it is this cognition that allows such a system to detect and understand the behavioural patterns of the human actors within the observed scene. Coupled with this is the requirement that the end user can communicate with the system to facilitate complex activities in the environment; this communication can be achieved either through context aware mobile devices that can adapt to dynamically changing environmental and physiological states or by external sensing and interpretation of the end user actions.

The driving goal of this research is to improve the efficiency, security and safety of airport based operations within the AmI paradigm. From a computer vision point of view this means the requirement of distributed visual surveillance and interpretation of a complex dynamic environment over extended time periods. In this paper we focus on the object tracking and data fusion modules from such a visual surveillance system; more details of the complete system are given in [5].

3 Scene Tracking

A motion detector segments an image into connected regions of foreground pixels, which is then used to track objects of interest across multiple frames. The motion detection algorithm selected for AVITRACK is the colour mean and variance algorithm (a background subtraction method based on the work of [3]). The evaluation process that led to this selection, is described in more detail in [6]. The colour mean and variance algorithm has a background model represented by a pixel-wise Gaussian distribution $N(\mu, \sigma^2)$ over the normalised RGB space, together with a shadow/highlight detection component based on the work of [7].

For per camera scene tracking, the feature-based KLT algorithm is incorporated into a higher-level tracking process to group features into meaningful objects; the individual features are subsequently used to associate objects to observations and to perform motion analysis when tracking objects during complex interactions.

For each object O, a set of sparse features S is maintained, with the number of features determined dynamically from the object's size and a configurable feature density parameter $\boldsymbol{\rho}$. The KLT tracker takes as input the set of observations $\{M_j\}$ identified by the motion detector, where M_j is a connected set of foreground pixels, with the addition of a nearest neighbour spatial filter of clustering radius r_c, i.e., connected components with gaps $\leq r_c$. Given such a set of observations $\left\{M_j^t\right\}$ at time t, and the set of tracked objects $\left\{O_i^{t-1}\right\}$ at $t-1$, object predictions $\left\{P_i^t\right\}$ are generated from the tracked objects. A prediction P_i^t is then associated with one or more observations, through a matching process that uses the individual tracking results of the features S of that object prediction and their spatial and/or motion information, in a rule-based approach.

The spatial rule-based reasoning method is based on the idea that if a feature belongs to object O_i at time $t-1$, then it should remain spatially within the foreground region of O_i at time t. A match function f is defined which returns the number of tracked features of prediction P_i^t that reside in the foreground region of observation M_j^t. The use of motion information in the matching process, is based on the idea that features belonging to an object should follow approximately the same motion (assuming rigid object motion). Affine motion models (solving for $w_t^T F w_{t-N} = 0$ [8]) are fitted to each group of k neighbouring features of P_i. These motion models are then represented as points in a motion parameter space and clustering is performed in this space to find the most significant motion(s) of the object. These motions are subsequently filtered temporally and matched per frame to allow tracking through merging/occlusion and identify splitting events.

On the apron, activity tends to happen in congested areas with several vehicles stationary in the proximity of the aircraft. To differentiate between stationary and moving objects, the motion detection process was extended to include a multi background layer technique. The tracker identifies stopped objects by performing region analysis of connected 'motion' pixels over a time window and by checking the individual motion of features of an object. Stationary objects are integrated into the motion detector's background model as different background layers. The advantage this method has over pixel level analysis (e.g. Collins *et al* [9]), is that for extended time periods (e.g. 30 minutes) pixel level methods tend to result in fragmented layers that do not represent cohesive objects.

To improve reasoning in the data fusion module we introduce a confidence measure that the 2-D measurement represents the whole object. The 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 ψ is estimated in an n pixel border of the scene as $\psi_e = \max(\left|\text{loc}_e(O_i) - \text{loc}_e(I_t)\right| / n, 1.0)$ where $e \in \{(\text{left}, x), (\text{bottom}, y), (\text{right}, x)\}$ determines for which edge of the image / object the confidence is measured, O_i is the object and I_t is the current image frame. ψ is in the range $0.0 - 1.0$, a single confidence estimate ψ_{O_i} is computed as a product over the processed bounding box edges for each object.

In the AVITRACK project both top-down and bottom-up approaches have been applied to the problem of object categorisation. The challenges faced in apron monitoring are the quantity (28 categories) and similarity of objects to be classified e.g. the majority of vehicles have similar appearance and size; therefore the simple descriptors used in many visual surveillance algorithms are likely to fail. The top-down approach [10, 2] applies a proven method to fit textured 3D models to the detected objects in the scene; the performance of this module is excellent for many of the vehicle categories with few false matches; the disadvantage of this method is the computational cost which is currently prohibitive. The bottom-up alternative to this approach is similar to the eigenwindow approach of Ohba and Ikeuchi [11]; this method has the advantage that objects can be classified even when partly occluded. The accuracy of the bottom-up method is currently 70% for limited classes of object. A more detailed description of the scene tracking process can be found in [12].

4 Data Fusion

The method applied for data fusion is based on a discrete nearest neighbour Kalman filter approach [1] with a constant velocity model; the main challenge in apron monitoring relates to the matching of tracks to observations; this is not solved by a probabilistic filter, therefore the simpler deterministic filter is sufficient. The (synchronised) cameras are spatially registered using coplanar calibration to define common 'world' co-ordinates. To localise objects in the world co-ordinates we devised a simple heuristic strategy that estimates the ground plane centroid using the camera angle to the ground plane, object category and the measured object size.

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. For each track, obtain the validated set of measurements per camera.
2. For each track, associate the nearest neighbour per camera.
3. Fuse associated measurements into a single measurement.
4. Kalman filter update of each track state with the fused measurement.
5. Inter-sensor association of remaining measurements to form candidate tracks.

The validated set of measurements are extracted using a validation gate [1]; this is applied to limit the potential matches between existing tracks and observations. In 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 – 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:

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

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:

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

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 coordinate system using the method presented in [13]. Measurement noise for the probability terms is determined *a priori*. An appropriate gate threshold can be determined from tables of the chi-square distribution [1].

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:

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

where $w_i = w_i' / \sum_{j=1}^{N} w_j'$ and $w_i' = 1/\psi_i^c$. ψ_i^c is the confidence of the i'th associated observation (made by camera c) estimated using the method in Section 3.

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. 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 [13]).

5 Experimental Results

The Motion Detection module is evaluated in previous work [6]. The Scene Tracking evaluation assesses the performance on representative test data containing challenging conditions for an objective evaluation. Two test sequences were chosen, Dataset 1 (2400 frames) contains the presence of fog whereas Dataset 2 (1200 frames) was acquired on a sunny day; both sequences contain typical apron scenes with congested areas containing multiple interacting objects.

The tracker detection rate ($TP/(TP{+}FN)$) and false alarm rate ($FP/(TP{+}FP)$) metrics defined by Black et al. [14] were used to characterise the overall tracking performance (where TP, FN and FP are the number of true positives, false negatives and false positives respectively). For Dataset 1 3435 true positives, 275 false positives and 536 false negatives were detected by the KLT based tracker. This leads to a tracker detection rate of 0.87 and a false alarm rate of 0.07. For Dataset 2 3021 true positives, 588 false positives and 108 false negatives were detected by the KLT based tracker. This leads to a tracker detection rate of 0.97 and a false alarm rate of 0.16. Representative results of the scene tracking module are presented in Figure 1. It can be seen that strong shadows are tracked as part of the mobile objects such as the tanker from Dataset 1 and the transporter from Dataset 2. In Dataset 1 a person (bottom-right of scene) leaves the ground power unit and in Dataset 2 a container is unloaded from the aircraft; these scenarios leave a ghost track in the previous object position.

The Data Fusion module is qualitatively evaluated for an extended sequence of Dataset 1 (9100 frames). The data fusion performance is shown in Figure 1 where estimated objects on the ground plane are shown; it can be seen that many of the estimated objects are contiguous. 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. Track identity can be lost when the object motion is not well modelled by the Kalman filter or when tracks are associated with spurious scene tracking measurements.

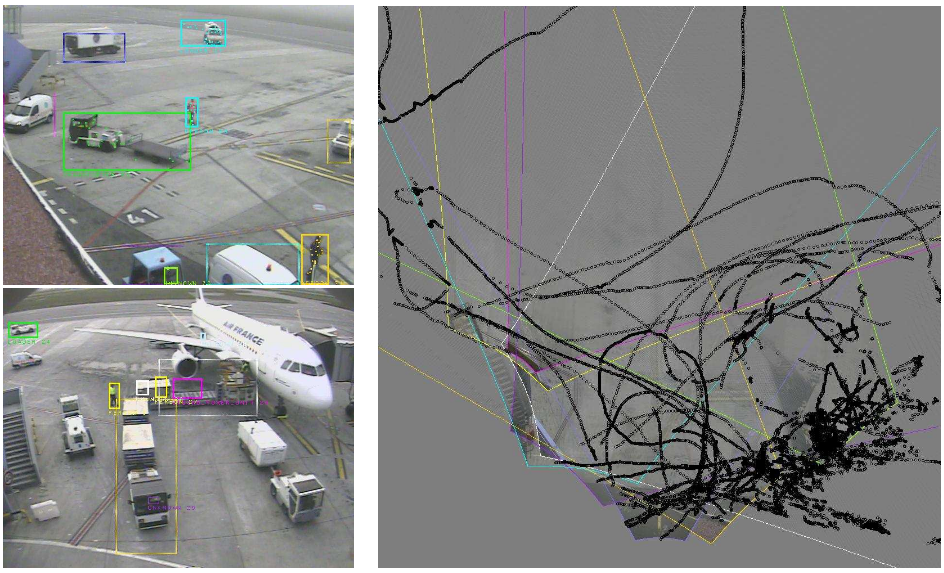

Fig. 1. (Left) Results obtained from the scene tracking module showing (Top) Dataset 1 and (Bottom) Dataset 2. (Right) Result obtained from the data fusion module.

6 Discussion and Future Work

The results are encouraging for both the Scene Tracking and Data Fusion modules; however, tracking is sensitive to significant dynamic and static object occlusions. Care must be taken to handle errors propagated from earlier modules, which can influence later processing stages (e.g. ghosts). Future work will look into using perspective projection motion models in the Scene Tracking module, speeding up the model based categorisation and using robust descriptors for the bottom-up method. In the Data Fusion module a particle filter based approach will be evaluated to improve performance in the presence of noise.

References

1. Bar-Shalom, Y., Li, X.: Multitarget-Multisensor Tracking: Principles and Techniques. YBS Publishing (1995)
2. Sullivan, G.D.: Visual interpretation of known objects in constrained scenes. In: Phil. Trans. R. Soc. Lon. Volume B, 337. (1992) 361–370
3. Wren, C.R., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: Real-time tracking of the human body. In: IEEE Transactions on PAMI. Volume 19 num 7. (1997) 780–785
4. Shi, J., Tomasi, C.: Good features to track. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition. (1994) 593–600
5. Thirde, D., Borg, M., Valentin, V., Fusier, F., Aguilera, J., Ferryman, J., Brémond, F., Thonnat, M., Kampel, M.: Visual surveillance for aircraft activity monitoring. In: Proc. Joint IEEE Int. Workshop on VS-PETS, Beijing. (2005)
6. Aguilera, J., Wildernauer, H., Kampel, M., Borg, M., Thirde, D., Ferryman, J.: Evaluation of motion segmentation quality for aircraft activity surveillances. In: Proc. Joint IEEE Int. Workshop on VS-PETS, Beijing. (2005)
7. Horprasert, T., Harwood, D., Davis, L.: A statistical approach for real-time robust background subtraction and shadow detection. In: IEEE ICCV'99 FRAME-RATE Workshop. (1999)
8. Xu, G., Zhang, Z.: Epipolar Geometry in Stereo, Motion and Object Recognition: A Unified Approach. Kluwer Academic Publ. (1996)
9. Collins, R., Lipton, A., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N., Hasegawa, O., Burt, P., Wixson, L.: A system for video surveillance and monitoring. In: Tech. Report CMU-RI-TR-00-12. (2002)
10. Ferryman, J.M., Worrall, A.D., Maybank, S.J.: Learning enhanced 3d models for vehicle tracking. In: Proc. of the British Machine Vision Conference. (1998)
11. Ohba, K., Ikeuchi, K.: Detectability, uniqueness, and reliability of eigen windows for stable verification of partially occluded objects. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. Volume 19 num 9. (1997) 1043–1048
12. Thirde, D., Borg, M., Aguilera, J., Ferryman, J., Baker, K., Kampel, M.: Evaluation of object tracking for aircraft activity surveillance. In: Proc. Joint IEEE Int. Workshop on VS-PETS, Beijing. (2005)
13. Black, J., Ellis, T.: Multi Camera Image Measurement and Correspondence. In: Measurement - Journal of the International Measurement Confederation. Volume 35 num 1. (2002) 61–71
14. Black, J., Ellis, T., Rosin, P.: A Novel Method for Video Tracking Performance Evaluation. In: Joint IEEE Int. Workshop on VS-PETS, Nice, France. (2003) 125–132