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

Visual Surveillance for Airport Monitoring Applications

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

[1]Pattern Recognition and Image Processing Group, Vienna University of Technology, Austria
agu@prip.tuwien.ac.at, kampel@prip.tuwien.ac.at

[2]Computational Vision Group, The University of Reading, UK
D.J.Thirde@reading.ac.uk, M.Borg@reading.ac.uk, J.Ferryman@reading.ac.uk

[3]Video & Safety Systems, ARC Seibersdorf research GmbH, Austria
Gustavo.Fernandez@arcs.ac.at

Abstract *This paper presents the Object Tracking component of a complete multi-camera surveillance system that was developed as part of the AVITRACK project. 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 overall Object Tracking component comprises three main modules: Motion detection, frame to frame object tracking and object categorisation.*

1 Introduction

This paper describes work undertaken on the EU project AVITRACK[1]. The main aim of this 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 multi-camera end-to-end system providing real-time recognition of the activities and interactions of numerous vehicles and personnel in a dynamic environment. The system described in this paper is realised at apron E-40 at Toulouse-Blagnac International Airport in France. The ability to recognise the apron activity on live real-time data is crucial to the success of the project and the complete system described in this paper is suitable for this task. Within this context we are focused on improving the degree of complexity that can be handled by existing real-time surveillance systems.

The tracking of moving objects on the apron has previously been performed using a top-down model based approach [16] although such methods are generally computationally expensive when applied to real-time tracking. 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.

Motion detection methods attempt to locate connected regions of pixels that represent the moving objects within the scene. Subsequent processes such as tracking and object classification are strongly dependent on them. There are many ways to detect moving objects including frame-to-frame differencing [11], background subtraction [8] and motion analysis [12] (e.g. optical flow) techniques. Background subtraction algorithms are the most common type of motion detectors. Such algorithms store an estimate of the static scene called background model, 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. The airport apron, being an outdoor environment, provides several challenges to motion detection. It must handle a wide range of environmental conditions, weather, and illumination changes, which can be long-term changes (diurnal cycle) or short-term (cloud movements, reflections, etc). The AVITRACK test sequences, like many CCTV applications, also suffer from chrominance and luminance sensitivity and have significant JPEG artifacts. The moving objects and apron are also of an achromatic nature with low contrast between the observed foreground and background.

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. Tracking algorithms have to deal with motion detection errors and complex object interactions; e.g. objects appear to merge together, occlude each other, fragment, undergo non-rigid motion, etc. Apron analysis presents further challenges due to the size of the vehicles tracked (e.g. the aircraft size is 34x38x12 metres), therefore prolonged occlusions occur frequently throughout apron operations. The apron can also be congested with objects; this enhances the difficulty of associating objects with regions. The Kanade-Lucas-Tomasi (KLT) [14] algorithm considers features to be independent entities and tracks each of them individually. The CamShift algorithm [4] uses appearance based (colour histogram) representation of objects to perform tracking using the mean-shift algorithm.

Object categorisation can be considered as the process of assigning class ownership to moving objects. Subsequent object behaviour analysis strongly depends on reliable ob-

[1] This work is supported by the EU, grant AVITRACK (AST3-CT-2003-502818).

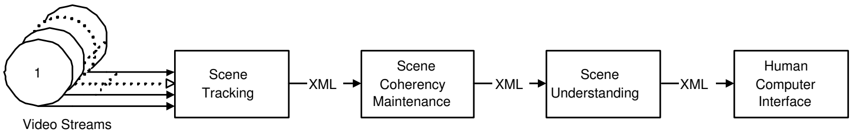


Figure 1: The AVITRACK System.

ject classification in order to have a firm basis upon which to act. This task is specially investigated by researchers since there is currently no method able to assure perfect classification of objects. Lipton et al. [11] propose the use of the dispersedness and the area of object blobs to classify the moving objects into vehicles, individuals and clutter. Kuno et al. [10] present a method that use simple shape parameters of human silhouette patterns to distinguish humans from other moving objects.

Most of the methods reviewed so far can be classified as bottom-up techniques, in that no high-level scene information is used in the classification process. Top-down approaches use high-level information about the scene to classify objects and recognise tasks. The work performed by [7] use both wire-frame (edge-based) 3D models as well as textured 3D models. 3D model-based classification provides high accurate results although the computation time required to process the models is high.

The remainder of this paper is organised as follows: Section 2 introduces the complete AVITRACK system. Section 3 discusses the object tracking module and Section 4 gives experimental results showing the performance of the proposed object tracking module.

2 The AVITRACK System

The system deployed is a decentralised multi-camera environment with overlapping fields of view (FOV); eight cameras are used in the operational prototype system to monitor the scene. This system is suitable for monitoring airport aprons since there are several camera mounting points on the airport building and overlapping fields of view are required to ensure consistent object labelling and enhanced occlusion reasoning within the scene. The majority of the mounting points observe the right hand side of the fuselage since this is where most of the servicing operations (such as baggage loading/unloading) take place; on the left hand side of the fuselage, the servicing operation of interest is the refuelling operation. 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.

2.1 System Architecture

The architecture of the system is shown in Figure 1 comprising four main processing modules - *Scene Tracking*, *Scene Coherency Maintenance*, *Scene Understanding*, and the *Human Computer Interface* module. The system also has additional offline modules that are used for camera calibration, a 2D/3D scene modelling module for generating geometric and semantic scene and object models, a video event definition module, and a module for replaying archived videos with annotated events.

A main requirement behind the adopted architecture is that the system must be capable of monitoring and recognising the activities and interaction of numerous vehicles and personnel in a dynamic environment over extended periods of time, operating in real-time at 12.5 FPS, with a frame resolution of 720×576 and using colour video streams.

Because of the relatively low quantity of distributed modules and the physical distances between them, the network operates via a standard 1Gb ethernet. The communications framework selected for the distributed modules is based on an OROCOS::SmartSoft CORBA [13] implementation. This allows for the use of naming services for the data broadcasted by a module, and for other modules to subscribe to it dynamically at run-time, thus allowing the network to be flexible and adapt at run-time to different hardware configurations.

The video streams are synchronised temporally and broadcasted by a central video server using the JPEG format. Initially the plan was for the high-volume video streams to be transmitted on dedicated connections; but after performing operational tests on the prototype system, it was found that the CORBA framework is fast enough to allow the video streams to be transmitted as CORBA services.

For the format of the data broadcasted by the other modules, the XML standard is used; although inefficient for communication over a network, the XML standard allows the system to be efficiently integrated as a series of black box modules with a defined interface between them. The partners in the project are able to develop the modules independently while adhering to the XML interface standard; this standardisation allowed the modules to be successfully integrated in the end-to-end system with few problems. The added advantage of the XML is that the human operators can manually inspect the XML to explain some system failures that may occur during integration.

2.2 Processing Modules

The main online processing modules comprising the AVITRACK system (see Figure 1) are:

- A *Scene Tracking* module comprising object tracking and data fusion sub-modules. The object tracker sub-module runs independently for each of the cameras, performing motion detection, frame to frame object tracking and object categorisation. Then, a central data fusion submodule receives the single-camera observations from the Frame Trackers, fuses the observations and generates 3D results to maximise the useful information content of the

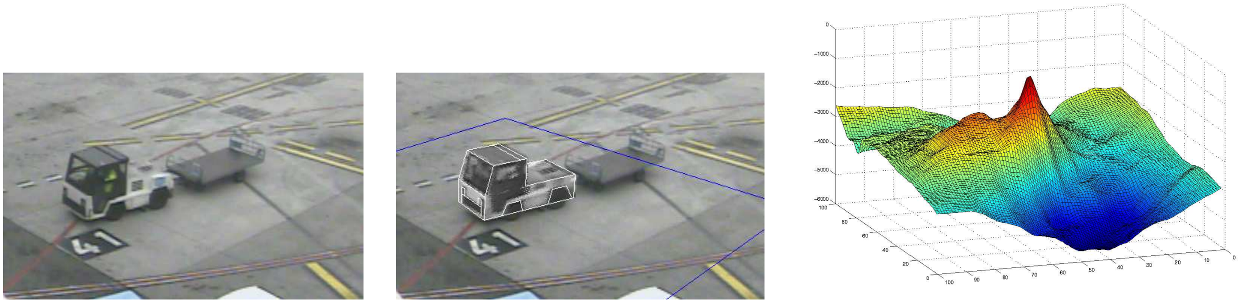


Figure 2: (Left) Frame of sequence S21 showing a transporter vehicle. (Centre) A 3D appearance model fitted to the vehicle, with the ground-plane (x,y) search area shown in blue. (Right) x,y-slice of the evaluation score surface in the $(x, y, \boldsymbol{\theta})$ search space.

observable scene.

- The *Scene Coherency Maintenance* module uses a temporal window and context information to improve the tracking results to provide enhanced trajectory information required for scene understanding.

- The *Scene Understanding* module uses the tracking results from the previous module to perform video event recognition and high level scene interpretation.

- The *Human Computer Interface* module is used to display the tracking results, the recognised events and any warnings/alarms associated with them as defined by the system operator.

In this paper the Object Tracker sub-module is described and evaluated. More details on the other processing modules can be found in [6, 17].

3 Object Tracking

The AVITRACK Object Tracker sub-module consists of motion detection (section 3.1) to find the moving objects in the observed scene, followed by object tracking in the image plane of the camera (section 3.2). The tracked objects are subsequently classified using a hierarchical object recognition scheme (section 3.3). In this Section we detail each step of the Object Tracker sub-module.

3.1 Motion Detection

The first task of the object 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. Three algorithms (all based on the aforementioned background subtraction method) were shortlisted in the evaluation process, as they were found to have acceptable susceptibility to noise and good detection sensitivity. These were the mixture of Gaussians [15], colour and edge fusion [9] and colour mean and variance [18]. After taking into account the evaluation results [1], the colour mean and variance method was the final choice for AVITRACK.

The colour mean and variance method models the background by a pixel-wise Gaussian distribution over the normalised RGB colour space. This algorithm was extended by a shadow/highlight detection component based on the work of Horprasert *et al* [8] to make it robust to illumination changes; and by using a multi-layered background approach to allow the temporary integration into the background of objects that become stationary for a short time period. More detail is given in [17].

3.2 Object Tracking

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. One such method, the Kanade-Lucas-Tomasi (KLT) feature tracker [14] 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. Features are considered to be independent entities which are tracked individually. Therefore, it 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, taking into account complex object interactions.

For each object O, a set of sparse local features S is maintained, with the number of features determined dynamically from the object's size and a feature density parameter. Using the observations (connected components of pixels) returned at time t by the motion detector, and the list of predictions generated from previously tracked objects (at $t-1$), the tracking process matches predictions to observations. The matching process uses both the spatial and the motion information of the local features. This is used within a rule-based framework that handles object merging and/or splitting events. The spatial rule-based reasoning is based on the idea that features are long-lived and if a feature belongs to an object O_i at time $t-1$, then the feature should remain spatially within the foreground region of O_i at time t; a function based on the spatial membership of features is used for matching. Motion information of features is also used, based on the idea that features belonging to an object should follow approximately the same motion. Affine motion models are fitted to the features of an object, which are then represented as points in a motion parameter space

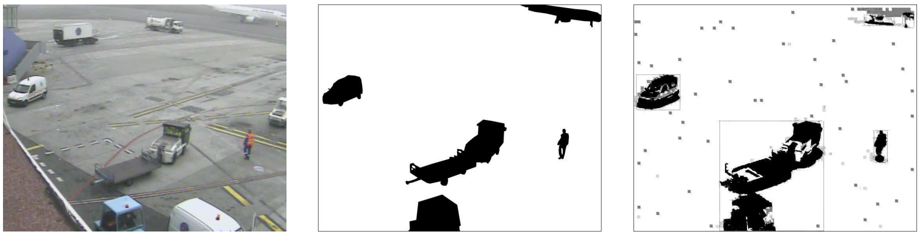


Figure 3: Representative motion detection result showing (Left) reference image, (Middle) ground truth and (Right) detection result.

and clustering is performed to find the most significant motion(s) of that object. These motions are filtered temporally and used to segment the observations at time t into distinct motions. More detail in [17].

3.3 Object Categorisation

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. After the first coarse classication, the second stage of the classication 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 [7] to fit textured 3D models to the detected objects in the scene.

Detailed 3D appearance models were constructed for the vehicles and encoded using the 'facet model' description language introduced in [16]. The model fit at a particular world point is evaluated by back-projecting the 3D model into the image and performing normalised cross-correlation (NCC) of the facet's appearance model with the corresponding image locations. To find the best fit for a model, the SIMPLEX algorithm is used to find the pose with best score in the search space, assuming the model's movements are constrained to be on the ground-plane. See Figure 2 for an example. More detail is given in [3].

4 Experimental Results

The Scene Tracking evaluation assesses the performance of the motion detection, the frame to frame object tracking and the categorisation components on representative test data.

4.1 Motion Detection Method

To evaluate the performance of the colour mean and variance motion detector six apron datasets were chosen. 20 reference frames were chosen from each dataset for which ground truth motion images were manually generated. These segmented objects were compared with the foreground objects detected by the motion detector. The performance evaluation of the colour mean and variance motion detector is described in more detail in [1].

Representative results of the colour mean and variance detector are depicted in Figure 3. It is noted that some objects are partially detected (See aircraft of Figure 3) due to the achromaticity of the scene and the presence of fog causes a relatively high number of foreground pixels to be misclassified as highlighted background pixels resulting in a decrease in accuracy. Weak shadows also cause problems, often detected as part of the mobile objects. Holes and fragmentation are presented in objects with the same colour as background.

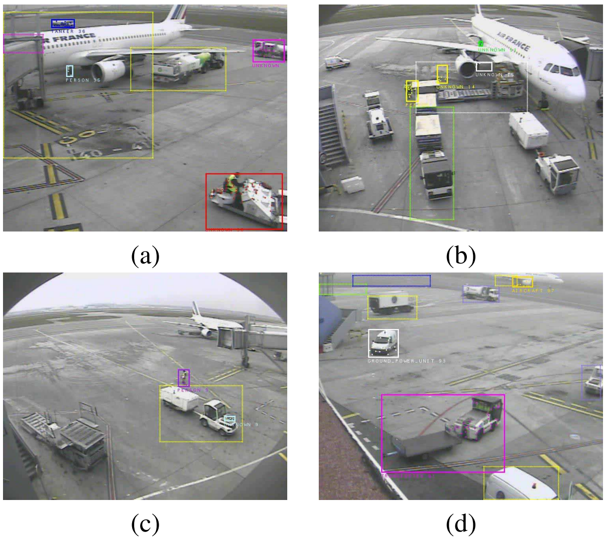
4.2 Local Feature Tracking Method

To evaluate the performance of the local feature tracking method four apron datasets were chosen. Dataset 1 (1100 frames), Dataset 2 (1100 frames) and Dataset 3 (1700) are acquired on a cloudy day whereas Dataset 4 (2200 frames) contains the presence of fog. The datasets have been manually annotated using ViPER annotation tool [5]. ViPER (Video Performance Evaluation Resource) is a semi-automatic framework designed to facilitate and accelerate the creation of ground truth image sequences and evaluate performance of algorithms. The ViPER's performance evaluation tool has been used to compare the result data of the local feature tracking method with the ground truth in order to generate data describing the success or failure of the performance analysis. At first, the evaluation tool attempts to match tracked objects (TO) to ground truth objects (GTO) counting objects as matches when the following metric distance is less than a given threshold.

$$D_i(t,g) = 1 - 2Area(t_i \wedge g_i)/(Area(t_i) + Area(g_i)) \quad (1)$$

Where t_i and g_i define the bounding-box of the tracked objects and ground truth objects at frame *i* respectively. Once the tracked and ground truth objects have been matched true positives (TP), false negatives (FN) and false positives objects (FP) are counted and summed up over the chosen frames. The metrics defined by Black et al. [2] were used to characterise the tracking performance.
Representative results of the local feature tracking method are presented in Figure 4.

Shadows are detected and tracked as part of the mobile objects such as the tanker from Dataset 1 and the transporter with containers from Dataset 2 (See Figure 4 (a, b)). In Figure 4 (b) a container is unloaded from the aircraft and in (c) cones are unloaded by a person from the front of the service vehicle. Both objects produce a ghost which remains behind the previous object position. An object is integrated into the background when becomes stationary. In these cases, ghosts are created when stationary objects start to move again. Furthermore, ghosts are produced when parts of the background start moving. Objects in the scene such as the aircraft from Figure 4 (d) are partially detected due to the achromaticity



(a) (b)

(c) (d)

Figure 4: The results obtained from the local feature based tracking algorithm. Image (a) has been chosen from Dataset 1, image (b) from Dataset 2 and images (c) and (d) from Dataset 3 and Dataset 4 respectively.

Dataset	TP	FP	FN	TRDR	FAR
1	2427	52	106	0.96	0.02
2	2413	92	392	0.86	0.04
3	2330	536	73	0.97	0.19
4	3724	69	218	0.94	0.02

Table 1: Performance results of the local feature tracking algorithm.

of the scene.

All ground truth objects of the evaluated datasets were matched to tracked objects. The tracker detection rate *TRDR* and the false alarm rate *FAR* were calculated for whole frames. The results of this evaluation are depicted in Table 1. The complexity of the scene presented in Dataset 2 which includes a high amount of occlusions causes a considerable number of false negatives provoking the decrease in *TRDR* (86%). Dataset 3 contains ghosts and reflections causing the increase in *FAR* (19%).

4.3 Object Categorisation Method

The evaluation of the object categorisation module was divided into two sub-tasks to reject the hierarchical method in which classification is performed: the per-frame bottom-up coarse-level classification for the main types of objects (people, vehicles, aircraft, equipment) and the detailed top-down vehicle recognition performed by 3D model-fitting in a background process. These are:

- Coarse categorisation: This task decides whether the object was correctly classified in the category or not.
- Recognition of the object in the category: When the object was correctly classified in its category, the object recognition task evaluate whether the category type of the object was correctly assigned or not.

Table 2 describes the possible categories of the objects in the evaluated datasets and for each category, the related subcategories are enumerated. The subcategories are necessary in order to differentiate objects with similar task or purpose (e.g. vehicles). For this evaluation, four sequences containing different object categories and subcategories were considered.

Category	Subcategories
Aircraft	Aircraft
Vehicle	GPU, tanker, transport and dollies, car, loader
Person	One person, group of people
Equipment	Container
Other	Other

Table 2: Category of the objects and correspondent subcategories.

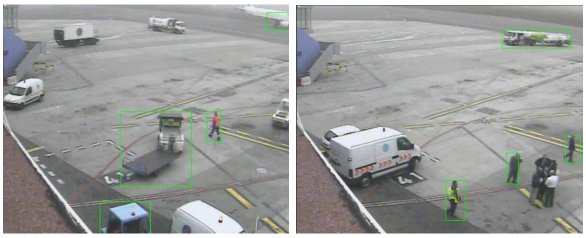


Figure 5: Object categorisation. a) Sequence 21, frame number 5842, categorised objects: Aircraft, two vehicles, and a person. b) Sequence 25, frame number 493, categorised objects: Vehicle and three people.

The evaluation procedure was done as follows: For each sequence, the evaluation was done frame by frame, checking whether objects present in the scene were properly classified into the appropriate category or not. At the same time, the recognition of the object by its subcategory was checked. When the classification of the object corresponds with the real type of the object, a true positive (TP) is counted. When the application assign an incorrect class to an object, a false positive (FP) is counted. Fig. 5 depicts categorisation results on different sequences. Table 3 summarises the categorisation results for each evaluated sequence, in terms of coarse-level and detailed level classification. It shows that some classification errors occur during the coarse-level classification. These errors appear especially in sequence 44 and sequence 10. The reason for this is that the bottom-up features used during the categorisation process are not properly detected, and therefore the categorisation process fails. But note the high accuracy obtained on the other two evaluated sequences. For the sub-type classification, more errors occur because of the similarity of several of the vehicles and also caused by incorrect model fitting by the SIMPLEX search algorithm (local minimum found instead of the global one).

5 Conclusion

The colour mean and variance motion detector has been tested on sequences with various weather conditions, including bright sunlight and fog. Gradual illumination changes can be handled without problems. Shadows provoke undesirable situations in 3D object reconstruction and tracking as

	Categorisation		Subcategorisation	
Sequence - Camera	T_+	F_+	T_+	F_+
10 - 8	73.77	26.23	68.89	31.11
21 - 7	97.86	2.14	77.38	22.62
22 - 5	91.03	8.97	61.31	38.69
44 - 4	60.13	39.87	88.93	11.07

Table 3: Object categorisation and object subcategorisation. Classification rates.

they are incorporated as part of the mobile objects. The low chromaticity information in the scene provokes detection errors reducing the sensitivity of the motion detector.

The evaluation of the local feature tracker demonstrates that the object tracking module detects a high proportion of the objects in the scene and these objects are tracked over extended time periods. Under severe partial partial occlusions the tracks become fragmented and lose the track ID.

Future work on the object tracker is to improve the prediction of the bounding boxes when object are undergoing occlusion and to retain the object ID's during this period. We also plan to work on the reducing the influencing of ghosts and reflections on the tracking procedure.

Considering object categorization, it is necessary to improve the categorisation method in order to avoid misclassification into the subcategories. A possible method to apply is to consider information of previous frames when an object has been previously detected and it was categorised. Such information might be possible to carry on from one frame to another like a history of the object's category.

References

[1] J. Aguilera, H. Wildenauer, M. Kampel, M. Borg, D. Thirde, and J. Ferryman. Evaluation of Motion Segmentation Quality for Aircraft Activity Surveillance. In *Proc. of the Second Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China*, pages 293–300, Oct. 2005.

[2] J. Black, T. Ellis, and P. Rosin. A Novel Method for Video Tracking Performance Evaluation. In *Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), Nice, France*, pages 125–132, 2003.

[3] M. Borg, D. Thirde, J. Ferryman, F. Fusier, V. Valentin, F. Bremond, and M. Thonnat. A real-time scene understanding system for airport apron monitoring. In *Proceedings of 2006 IEEE International Conference on Computer Vision Systems*, New York, USA, Jan 5-7, 2006. IEEE Computer Society.

[4] G. Bradski. Computer Vision Face Tracking for Use in a Perceptual User Interface. In *Intel Technology Journal*, volume Q2, 1998.

[5] D. Doermann and D. Mihalcik. Tools and Techniques for Video Performance Evaluation. In *Proceedings of the International Conference on Pattern Recognition, Barcelona, Spain*, volume 4, pages 167–170, September 2000.

[6] J. Ferryman, M. Borg, D. Thirde, F. Fusier, V. Valentin, F. Bremond, M. Thonnat, J. Aguilera, and M. Kampel. Automated Scene Understanding for Airport Aprons. In *Proc. of the Australian Joint Conference on Artificial Intelligence, Sydney, Australia*, pages 593–603, Dec. 2005.

[7] J.M Ferryman, A.D. Worral, and S.J Maybank. Learning Enhanced 3D Models for Vehicle Tracking. In *Proceedings of the Bristish Machine Vision Conference, Southampton, UK*, pages 873–882, 1998.

[8] Thanarat Horprasert, David Harwood, and Larry S. Davis. A Statistical Approach for Real-Time Robust Background Subtraction and Shadow Detection. In *Proceedings of IEEE ICCV FRAME-RATE Workshop, Kerkyra, Greece*, pages 1–19, September 1999.

[9] S. Jabri, Z. Duric, H. Wechsler, and A. Rosenfeld. Detection and Location of People in Video Images Using Adaptive Fusion of Color and Edge Information. In *Proceedings of the International Conference on Pattern Recognition*, pages 4627–4631, 2000.

[10] Y. Kuno, T. Watanabe, Y. Shimosakoda, and S. Nakagawa. Automated Detection of Human for Visual Surveillance System. In *Proc. of the International Conference on Pattern Recognition*, volume 24 Issue 4, pages 509–522, 2002.

[11] A.J. Lipton, H. Fujiyoshi, and R.S. Patil. Moving Target Classification and Tracking from Real-Time Video. In *Proc. IEEE Workshop Applications of Computer Vision*, pages 8–14, 1998.

[12] D. Meyer, J. Denzler, and H. Niemann. Model Based Extraction of Articulated Objects in Image Sequences for Gait Analysis. In *Proc. IEEE Int. Conf. Image Processing*, pages 78–81, 1998.

[13] C. Schlegel. A Component Approach for Robotics Software: Communication Patterns in the Orocos Context. In *Proc. of the Fachtagung Autonome Mobile Systeme (AMS), Informatik aktuell, Springer, Karlsruhe*, pages 253–263, 2003.

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

[15] C. Stauffer and W.E.L Grimson. Adaptive Background Mixture Models for Real-Time Tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 246–252, 1999.

[16] G. D. Sullivan. Visual Interpretation of Known Objects in Constrained Scenes. In *Phil. Trans. R. Soc. Lon.*, volume B, 337, pages 361–370, 1992.

[17] D. Thirde, M. Borg, V. Valentin, F. Fusier, J. Aguilera, J. Ferryman, F. Bremond, M. Thonnat, and M. Kampel. Visual Surveillance for Aircraft Activity Monitoring. In *Proc. of the Second Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China*, pages 255–262, Oct. 2005.

[18] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-Time Tracking of the Human Body. In *IEEE Transactions on PAMI*, volume 19 num 7, pages 780–785, 1997.