

The AIT Outdoor Tracker for Vehicles and Pedestrians in CLEAR2007

Andreas Stergiou, Aristodemos Pnevmatikakis and Lazaros Polymenakos

Athens Information Technology, Autonomic and Grid Computing,
P.O. Box 64, Markopoulou Ave., 19002 Peania, Greece
{aste, apne, lcp}@ait.edu.gr
<http://www.ait.edu.gr/research/RG1/overview.asp>

Abstract. This paper presents the tracking system from Athens Information Technology that participated to the pedestrian and vehicle surveillance task of the CLEAR 2007 evaluations and the obtained results. The system is based on the CLEAR 2006 one, with some important modifications that are detailed. Since the test data in CLEAR 2006 and 2007 are the same, it is easy to quantify the obtained performance gain from the older system to the proposed one.

1 Introduction

Target tracking in video streams has many applications, like surveillance, security, smart spaces [1], pervasive computing, and human-machine interfaces [2] to name a few. In these applications the objects to be tracked are either humans, or vehicles. To track objects we first need to detect them. The detected objects are used to initialize the tracks and provide measurements to the tracking algorithm, usually of the recursive Bayesian filtering [3] type. This is a very hard problem, one that remains unsolved in the general case [3]. If a shape or a color model of the objects were known a-priori, then detection can be done using active contours [4] or variations of the mean-shift algorithm [5]. Unfortunately such approaches can only be applied in limited application domains; the shape and color richness of all possible people and vehicles prohibit their use in unconstrained applications like surveillance or smart rooms.

The solution to the detection problem is a common property of such targets: sooner or later they move, which produces evidence that distinguishes them from the background and identifies them as foreground objects. The segmentation of foreground objects can be accomplished by processing the difference of the current frame from a background image. This background image can be static [6] or can be computed adaptively [7]. The drawback of the static background image is that background does change. In outdoor scenes natural light changes and the wind causes movement of trees and other objects. In indoor scenes, artificial light flickers and pieces of furniture may be moved around. All such effects can be learned by an adaptive background algorithm like Stauffer's [8] and of its modifications [9,10]. Such an algorithm detects targets as segments different from the learned background,

but depends on the targets' movement to keep a fix on them. If they stop, the learning process fades them into the background.

Once a target is initialized, a tracking system should be able to keep a fix on it even when it remains immobile for some time. In this paper, we propose a novel tracking system that addresses this need by utilizing a feedback mechanism from the tracking module to the adaptive background module which in turn provides the evidence for each target to the tracking module. We control the adaptive background parameters on a pixel level for every frame (spatiotemporal adaptation), based on a prediction of the position of the target. Under the assumption of Gaussian target states and linear dynamic models, this prediction can be provided by a Kalman filter [11].

The proposed tracker system comprises of the feedback configuration of three modules, namely the adaptive background, the image processing for evidence generation and the Kalman filtering. A fourth module operates on the tracks in a temporal window of 1 second, by checking their consistency.

This paper is organized as follows: In section 2 the four modules of the system are detailed. The results on the VACE person and vehicle surveillance tasks of the CLEAR 2006 evaluations are presented and discussed in section 3. Finally, in section 4 the conclusions are drawn, followed by some indications for future enhancements.

2 Tracking System

The block diagram of the tracking system is shown in Figure 1. It comprises four modules: adaptive background, image processing for evidence generation, Kalman filtering and track consistency. Evidence for the targets is generated once difference from the estimated background is detected. The estimation of the background is dynamic; background is learnt in a different manner for different portions of the frame, depending on whether they belong to existing targets, the target size and its speed. The dynamic estimation in the CLEAR 2007 system is applied on a blurred version of the original frame, to reduce the effect of flicker. The evidence is used to initialize and update tracks. Tracks that are persistent for 10 out of the 15 past frames are promoted to targets, and are reported by the system. Given that the frame rate of all the VACE person and vehicle surveillance videos is 25 per second, the introduced lag is a small penalty to pay for the added robustness to false alarms. Initialized tracks have their new position predicted by the state update step of the Kalman filter. The predictions are used to associate evidence with tracks and perform the measurement update step of the Kalman filter. Tracks are also eliminated if they have no evidence supporting them for 15 frames. The states of the Kalman filter, i.e. the position, velocity and size of the targets, are fed back to the adaptive background module to spatiotemporally adapt the learning rate. They are also fed forward to the track consistency module to obtain the reported tracks of the system and the decision whether they correspond to vehicles or pedestrians. In the rest of the section, we present the four modules in detail.

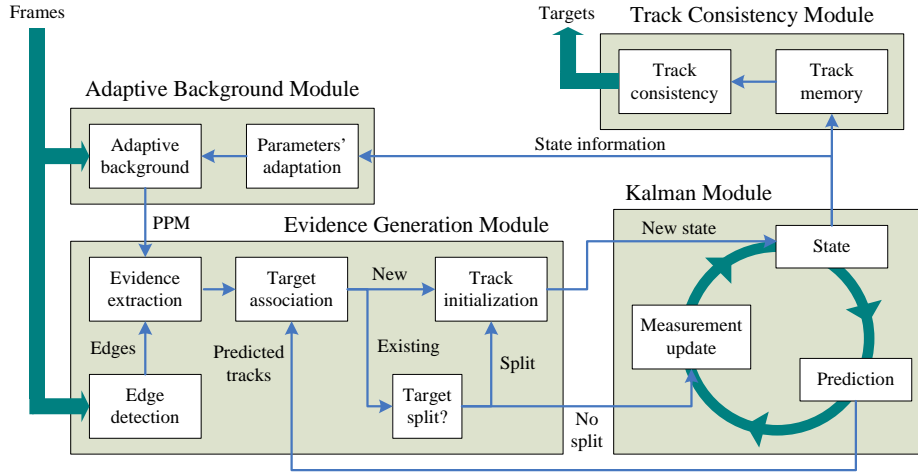


Fig. 1. Block diagram of the complete feedback tracker architecture. Frames are input to the adaptive background and evidence generation modules, and targets are output from the track consistency module.

2.1 Adaptive Background Module

The targets of the proposed system (vehicles and pedestrians) are mostly moving. The changes in subsequent video frames due to movement are used to identify and segment the foreground (pixels of the moving targets) from the background (pixels without movement). If a background image were available, this segmentation is simply the difference of the current frame from the background image. The foreground pixels thus obtained are readily grouped into target regions. A static image of the empty scene viewed by the (fixed) camera can be used for background [6]. Unfortunately this is not practical for outdoors applications, or even for long term indoors applications, hence adaptive background approaches are adopted [7-10] primarily for two reasons: First, such an empty scene image might not be available due to system setup. Second and most important, background (outdoors and indoors) also changes: Natural light conditions change slowly as time goes by; the wind causes swaying movements of flexible background object (e.g. foliage); fluorescent light flickers at the power supply frequency; objects on tabletops and small pieces of furniture are rearranged and projection areas display different content. All these changes need to be learnt into an adaptive background model.

Stauffer's adaptive background algorithm [8] is capable of learning such changes with different speeds of change by learning into the background any pixel, whose color in the current frame resembles the colors that this pixel often had in the history of the recording. So no changes, periodic changes or changes that occurred in the distant past lead to pixels that are considered background. To do so, a number of weighted Gaussians model the appearance of different colors in each pixel. The weights indicate the amount of time the modeled color is active in that particular

pixel. The mean is a three dimensional vector indicating the estimated color for that model and that pixel, while the covariance matrix indicates the extend around the mean that a color of that pixel is to be considered similar to the one modeled. Colors in any given pixel similar to that modeled by any of the Gaussians of that pixel lead to an update of that Gaussian, an increase of its weight and a decrease of all the weights of the other Gaussians of that pixel. Colors not matching any of the Gaussians of that pixel lead to the introduction of a new Gaussian with minimum weight. Hence the possible updates of the weight of the i -th Gaussian of the pixel located at (x, y) at time t are

$$w_i(x, y, t) = \begin{cases} a & \text{new Gaussian} \\ (1-a)w_i(x, y, t-1) & \text{non-matching Gaussians} \\ (1-a)w_i(x, y, t-1) + a & \text{matching Gaussians} \end{cases} \quad (1)$$

where a is the learning rate.

Some variations of the Stauffer algorithm found in the literature deal with the way covariance is represented (single value, diagonal of full matrix) and the way the mean and covariance of the Gaussians are updated [9]. Some further variations of the algorithm address the way the foreground information is represented. The original algorithm and most of the modifications lead to a binary decision for each pixel: foreground or background [8,9]. In [10], the Pixel Persistence Map (PPM) is used instead. This is a map of the same dimension as the frames with a value at each location (x, y) equal to the weight of the Gaussian matching the current color of the pixel at (x, y) . Small PPM values indicate foreground objects, while large indicate background. The foreground/background threshold is left unspecified though.

The drawback of all the existing variations of Stauffer's algorithm is that stationary foreground objects tend to fade in the background with rate a . Small background learning rates fade foreground objects slowly, but are also slow in adapting to the background changes. Large rates favor background adaptation but tend to fade a target into the background when it stops. This fading progressively destroys the region of the tracked object, deforms its perceived shape and finally leads to loosing track of the object altogether. When the target resumes moving, foreground pixels will be marked only at the locations not previously occupied by the stationary target. When a target remains stationary long enough, or has fairly uniform coloration, the new evidence will be far apart from the last evidence of the track, either in time, or in space or in both. Then the track is lost; the track is terminated and another is initiated when movement resumes.

We address the problem of the fading of stationary foreground objects using a feedback tracking architecture. The edges of the frame that coincide with values of the PPM below a threshold serve as target evidence to the Kalman filter. The states of the Kalman filter provide ellipses that describe every target. The learning rate is modified in regions around these targets, based on their speed and size. Thus, instead of a constant value, a spatiotemporal adaptation of the learning rate is used:

$$a(x, y, t) = \begin{cases} 0.04 & \text{if } (x, y) \text{ not a target pixel at time } t \\ a(\hat{v}, C) & \text{if } (x, y) \text{ a target pixel at time } t \end{cases} \quad (2)$$

where C is the covariance matrix of the target (hence $\det(C)$ relates to its size) and \hat{v} is the mobility of the target, which is related to the change of the position of its centroid and the change of its size. The latter indicates an approaching or receding target and is quantified using the determinant of the covariance matrix of the target. Thus the mobility is defined as follows:

$$\hat{v} = T_f \sqrt{\|\mathbf{v}\|_2} + \sqrt{\frac{\max(C_t, C_{t-T_f})}{\min(C_t, C_{t-T_f})}} \quad (3)$$

where \mathbf{v} is the velocity vector and T_f is the inverse of the frame rate. Then, the learning rate $a(\hat{v}, C)$ of a pixel belonging to a target is:

$$a(\hat{v}, C) = \begin{cases} 0.04 & \text{if } \det(C) \leq 8 \cdot 10^5 \text{ and } \hat{v} \geq 2 \\ \frac{0.04}{4} & \text{if } \det(C) > 8 \cdot 10^5 \text{ and } \hat{v} \geq 2 \\ 0.0044 \cdot \tan\left(\frac{\hat{v} \cdot \pi}{4.3}\right) & \text{if } \det(C) \leq 8 \cdot 10^5 \text{ and } \hat{v} < 2 \\ \frac{0.0044}{4} \cdot \tan\left(\frac{\hat{v} \cdot \pi}{4.3}\right) & \text{if } \det(C) > 8 \cdot 10^5 \text{ and } \hat{v} < 2 \end{cases} \quad (4)$$

This choice for $a(\hat{v}, C)$ progressively delays fading of the targets as they become slower. It also delays fading of large targets by setting the learning rate to $1/4$ of its value if the target is too large. This is useful for large vehicles, where their speed can be large, but their uniform colors can lead to fading into the background.

The second major proposed modification of Stauffer's algorithm addresses extreme flickering situations often encountered in night vision cameras. In such scenes the PPM needs to be binarized by a high threshold in order not to consider flickering pixels as foreground. The high threshold on the other hand tends to discard actual foreground pixels as well. The proposed solution is to adapt the threshold T in a spatiotemporal fashion similar to the learning rate in (2). i.e.

$$T(x, y, t) = \begin{cases} 0.25 & \text{if } (x, y) \text{ not a target pixel at time } t \\ & \text{or a target with } \det(C) < 500 \\ 0.5 & \text{elsewhere} \end{cases} \quad (5)$$

This way flickering pixels are avoided far from the targets, while the targets themselves are not affected. To avoid a delayed detection of new very small targets, the threshold of pixels belonging to such targets with $\det(C) < 500$ is not affected.

These proposed feedback mechanisms on the learning rate and PPM binarization threshold lead to robust foreground regions regardless of the flickering in the images or the lack of target mobility, while they do not affect the adaptation of the background around the targets. When such flickering and mobility conditions occur,

the resulting PPM is more suitable for target region forming than the original version of [10]. The forming of target regions is the goal of the evidence generation module, detailed next.

2.2 Evidence Generation Module

The evidence generation module finds foreground segments, assigns their pixels to known targets or initializes new ones and checks targets for possible splitting. The information for new targets or targets to be updated is passed to the Kalman module.

The binary mask obtained by adaptively thresholding the PPM is passed through a shadow detector based on [12]. It is then merged with the binary mask obtained by the edge detector using the AND operator. The resulting mask contains the foreground edges. Edges are used to add robustness to the system: The PPM can have filled segments if the object has entered the camera view after initialization and moves sufficiently. On the other hand some object can manifest itself by parts of its outline if it has been present in the scene at initialization. The use of edges provides the contours of objects in both cases, so they no longer need to be treated by different image processing modules. The foreground edge map is dilated to form regions. The dilation is such that edges that lie up to 10 pixels apart are merged into a single segment. Should these regions contain holes, they are filled, providing solid regions. This modification of the CLEAR 2007 system compared to the CLEAR 2006 one is important, as it results to robust estimation of the Kalman filter states.

The association of the evidence to targets is the major modification of the CLEAR 2007 system. While the CLEAR 2006 association was done on a per segment basis, it is performed on a per pixel basis in the CLEAR 2007 system. The advantage of this per pixel approach is robustness to:

- target splits, when the evidence of a target is split into more than one segment due to similarity to the background color, and
- collision handling, when targets touch and their evidences produce one big segment that can be shared between multiple targets.

The states of the targets contain their mean and the covariance matrix, hence the Mahalanobis distance of any evidence pixel from a target can be found. The association of the evidence to the targets utilizes this Mahalanobis distance as follows:

- Allocate to every existing target all evidence pixels that have Mahalanobis distance smaller than 1. By doing so pixels can be allocated to more than one target, while others can be left unallocated.
- For every unallocated evidence pixel, find the closest target. If the distance is less than 2, associate it to the target. Each of these pixels is hence either allocated to a single target, or it still remains unallocated.
- If a distinct region of evidence pixels has none of its member pixels allocated to a target, then this region is used to initialize a new target.

After the association of evidence pixels to targets, the regions of the targets are checked for possible split. A split is allowed if the target is big enough (in terms of number of pixels) and its modeling with a two-dimensional Gaussian density is not good enough. The fitness of the two-dimensional Gaussian model is enumerated by

the total probability of the associated pixels. If this falls below 0.5, the target is split in two using k-means.

2.3 Kalman Filtering Module

The Kalman filtering module maintains the states of the targets. It creates new targets should it receive a request from the evidence generation module and performs measurement update based on the foreground segments associated to the targets. The states of the targets are fed back to the adaptive background module to adapt the learning rate and the threshold for the PPM binarization. States are also eliminated if they have no foreground segments associated to them for 15 frames.

Every target is approximated by an elliptical disc that is obtained by the mean \mathbf{m} and the covariance matrix \mathbf{C} of the target, i.e. it is described by a single Gaussian.

$$\begin{aligned} \mathbf{m} &= [m_x, m_y]^T \\ \mathbf{C} &= \begin{bmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{bmatrix} \end{aligned} \quad (6)$$

If the eigenvectors and the eigenvalues of \mathbf{C} are \mathbf{v}_i and λ_i respectively, with $i = 1, 2$, then the axes of the ellipse are the \mathbf{v}_i and the radii are $2\sqrt{\lambda_i}$.

The target states are seven-dimensional; they comprise of the mean of the Gaussian describing the target (horizontal and vertical components), the velocity of the mean (horizontal and vertical components) and the three independent terms of the covariance matrix. Hence the state vector is:

$$\mathbf{s} = [m_x, m_y, v_x, v_y, C_{11}, C_{22}, C_{12}]^T \quad (7)$$

The prediction step uses a loose linear dynamic model of constant velocity [17] for the update of the mean position and velocity. As for the update of the three covariance terms, their exact model is non-linear, hence cannot be used with the Kalman tracker; instead of using linearization and an extended Kalman tracker, the covariance terms are modeled as constant. The variations of the velocity and the covariance terms are permitted by the state update variance term. This loose dynamic model permits arbitrary movement of the targets. It is very different to the more elaborate models used for tracking aircraft. Aircraft can perform a limited set of maneuvers that can be learned and be expected by the tracking system. Further, flying aircraft can be modeled as rigid bodies thus strict and multiple dynamic models are appropriate and have been used extensively in Interacting Multiple Model Kalman trackers [18,19]. Unlike aircraft, street vehicles and especially humans have more degrees of freedom for their movement which includes apart from speed and direction changes obstacles arbitrarily, rendering the learning of a strict dynamic model impractical. A strict dynamic model in this case can mislead a tracker to a particular track even in the presence of contradicting evidence [3].

2.4 Track Consistency Module

The track consistency module has two roles: Firstly, it manages the tracks for possible elimination and merging. Secondly it classification the targets into pedestrians and vehicles. The input to the track management is the states of the Kalman filtering module. The processed tracks are the input to the target type classification.

Track management begins with elimination of tracks when they are:

- very short (in duration),
- not spanning any distance (immobile objects are mostly false positives), and
- mainly existing inside other targets (such targets are usually vehicle fragments inside a larger vehicle bounding box).

Then, adjacent tracks are merged, to avoid identity switches. To do so, the starting and ending points of tracks are considered in time-space. If the temporal proximity is less than 0.6 seconds (15 frames) and the spatial proximity is less than 30 pixels, then the tracks are merged.

The decision about the type of target is based on the velocity, size and the distance covered by the target. A vehicle confidence is built using the product of these metrics. If the product is large enough, then the target is considered a vehicle. This approach fails when vehicle tracks are fragmented by some occlusion. Then some too short (in space) track can lead to wrong classification.

3 CLEAR 2007 Evaluation Results

The outdoor videos that the CLEAR 2007 system is tested on are the same as those of CLEAR 2006. Figure 2 contains typical frames from these recordings, with the detected targets and their tracks superimposed. The recordings are 50 in total depicting two different sites at different whether conditions. The main difference between the two sites is the motion of the vehicles; in site 2 they park/un-park performing slow and jerky maneuvers. In site 1 there are some videos with adverse whether conditions: wind is moving the background a lot, and possibly the camera a bit, while very bright sunlight casts long shadows. For this reason the results are grouped in these three categories.

The quantitative evaluation of the system is based on the metrics described in [20]. The primary metrics for face tracking are two: The Multiple Object Tracking Precision (MOTP) is the position error for all correctly tracked targets over all frames. It is a measure of how well the system performs when it actually finds the target. There are three kinds of errors for the tracker, false positives, misses and track identity mismatches. They are reported jointly in an accuracy metric, the Multiple Object Tracking Accuracy (MOTA). The MOTA is the residual of the sum of these three error rates from unity. The mean MOTA of the CLEAR 2007 system is compared to that of the CLEAR 2006 in Table 1. Table 2 summarizes the MOTP.



Fig. 2. Typical frames from the outdoor recordings with the tracked targets superimposed. Dashed bounding boxes and lines correspond to vehicles and their tracks.

Evidently the MOTA has improved overall and per site, conditions and target type. The only exception is the adverse conditions in site 1, where the MOTA is a bit decreased for pedestrians only. The most impressive performance gains can be found in site 2 for pedestrians and site 1 (adverse conditions) and site 2 for vehicles. The CLEAR 2007 system for vehicle tracing performs similarly across sites and conditions. The same system for pedestrians is still short on performance for site 2. Excluding site 2 for pedestrians, the system tracks equally well pedestrians and vehicles.

Table 1. Mean MOTA of the CLEAR 2007 tracking system, compared to that of 2006 for the different sites and conditions.

Recordings	Pedestrians		Vehicles	
	2006	2007	2006	2007
All	0.00277	0.255	0.312	0.511
Site 1, normal	0.526	0.574	0.517	0.556
Site 2	-0.906	-0.261	0	0.456
Site 1, adverse	0.489	0.462	0.242	0.477

Regarding tracking precision, the MOTP of the CLEAR 2007 system for pedestrians has dropped compared to the 2006 one. This is mainly due to the inclusion of more targets, which due to their location are harder to track. Regarding vehicles, precision has increased, indicating a better framing of vehicles by the new system. The MOTA is consistent across sites and target types, with the only exception of the

adverse conditions in site 1. The long shadows caused by the bright sunlight in these recordings result to looser framing of the pedestrians.

Table 2. Mean MOTP of the CLEAR 2007 tracking system, compared to that of 2006 for the different sites and conditions.

Recordings	Pedestrians		Vehicles	
	2006	2007	2006	2007
All	0.525	0.434	0.385	0.434
Site 1, normal	0.556	0.429	0.561	0.447
Site 2	0.461	0.456	0	0.412
Site 1, adverse	0.577	0.395	0.491	0.432

The spread of the MOTA scores is analyzed by the boxplot of the MOTA per target type depicted in Figure 3. Evidently the spread is reduced in the CLEAR 2007 system compared to the 2006 one.

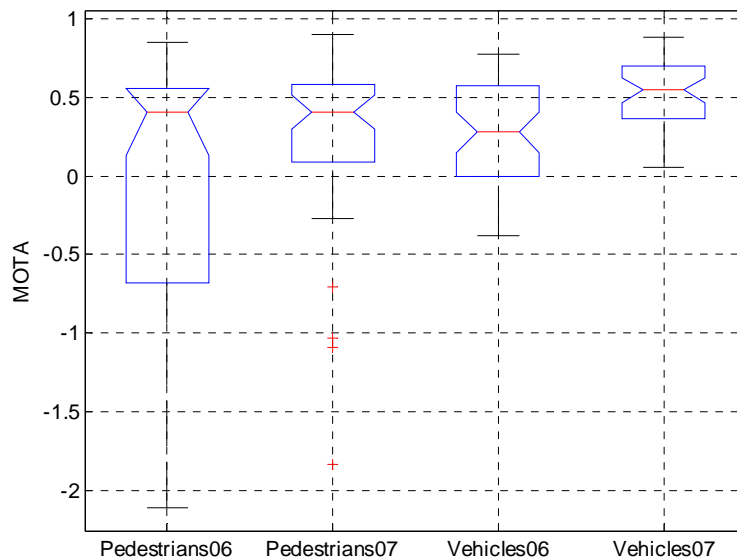


Fig. 3. Boxplot comparison of the MOTA per target type in the CLEAR 2006 and 2007 evaluations. The improvement of the median value of the MOTA for pedestrians is not statistically significant, even though the newer system has negative MOTA only for the outliers. The improvement of the median value of the MOTA for vehicles is statistically significant.

4 Conclusions

The AIT system for pedestrian and vehicle tracking for CLEAR 2007 evaluations has superior performance to that for the CLEAR 2006. The most important modifications are the way the evidence is associated to the tracked targets, and the track consistency module that manages the tracks and classifies targets into pedestrians and vehicles.

A number of issues remain open. Firstly, the degraded performance on pedestrians for the site 2 recordings needs to be analyzed. To do so, the MOTA score needs to be broken down to its components: the miss, false positive and identity switch rates. Secondly, color information can be incorporated in the evidence generation. Finally, a tracker based on particle filters and the CONDENSATION [21] algorithm can be incorporated in the feedback loop of the filter.

Acknowledgements

This work is sponsored by the European Union under the integrated project CHIL, contract number 506909. The authors wish to thank the people involved in data collection, annotation and overall organization of the CLEAR 2007 evaluations for providing such a rich test-bed for the presented algorithm.

References

- [1] A. Waibel, H. Steusloff, R. Stiefelhagen, et. al: CHIL: Computers in the Human Interaction Loop, *5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, Lisbon, Portugal, (Apr. 2004).
- [2] A. Pnevmatikakis, F. Talantzis, J. Soldatos and L. Polymenakos: Robust Multimodal Audio-Visual Processing for Advanced Context Awareness in Smart Spaces, in I. Maglogiannis, K. Karpouzis and M. Bramer (eds.), *Artificial Intelligence Applications and Innovations (AIAI06)*, Springer, Berlin Heidelberg (June 2006), 290-301.
- [3] D. Forsyth and J. Ponce: *Computer Vision - A Modern Approach*, Prentice Hall, (2002), 489-541.
- [4] J. MacCormick: Probabilistic modelling and stochastic algorithms for visual localisation and tracking, PhD Thesis, University of Oxford (2000), section 4.6.
- [5] G. Jaffré and A. Crouzil: Non-rigid object localization from color model using mean shift, *International Conference on Image Processing (ICIP 2003)*, Barcelona, Spain, (Sept. 2003)
- [6] H. Ekenel and A. Pnevmatikakis: Video-Based Face Recognition Evaluation in the CHIL Project – Run 1, *Face and Gesture Recognition*, Southampton, UK, (Mar. 2006), 85-90.
- [7] A. McIvor: Background Subtraction Techniques, *Image and Vision Computing New Zealand*, (2000).
- [8] C. Stauffer and W. E. L. Grimson: Learning patterns of activity using real-time tracking, *IEEE Trans. on Pattern Anal. and Machine Intel.*, 22, 8 (2000), 747–757.
- [9] P. KaewTraKulPong and R. Bowden: An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection, in *Proc. 2nd European Workshop on Advanced Video Based Surveillance Systems (AVBS01)*, (Sept 2001).

- [10] J. L. Landabaso and M. Pardas: Foreground regions extraction and characterization towards real-time object tracking, in *Proceedings of Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI '05)*, (July 2005).
- [11] R. E. Kalman: A New Approach to Linear Filtering and Prediction Problems, *Transactions of the ASME – Journal of Basic Engineering*, 82 (Series D), (1960) 35-45.
- [12] L.-Q. Xu, J. L. Landabaso and M. Pardas: Shadow Removal with Blob-Based Morphological Reconstruction for Error Correction, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, (March 2005).
- [13] S. Blackman: Multiple-Target Tracking with Radar Applications, Artech House, Dedham, MA, (1986), chapter 14.
- [14] Z. Zhang: A Flexible New Technique for Camera Calibration, Microsoft Research, Technical Report MSR-TR-98-71, (Aug. 2002).
- [15] M. Jones and J. Rehg: Statistical color models with application to skin detection, *Computer Vision and Pattern Recognition*, (1999), 274–280.
- [16] P. Viola and M. Jones: Rapid Object Detection using a Boosted Cascade of Simple Features, *IEEE Conf. on Computer Vision and Pattern Recognition*, (2001).
- [17] S.-M. Herman: A particle filtering approach to joint passive radar tracking and target classification, PhD thesis, University of Illinois at Urbana-Champaign, (2002), 51-54.
- [18] H. A. P. Bloom and Y. Bar-Shalom: The interactive multiple model algorithm for systems with Markovian switching coefficients, *IEEE Trans. Automatic Control*, 33 (Aug. 1988), 780-783.
- [19] G. A. Watson and W. D. Blair: IMM algorithm for tracking targets that maneuver through coordinated turns, in *Proc. of SPIE Signal and Data Processing of Small Targets*, 1698 (1992), 236-247.
- [20] R. Kasturi, et. al: Performance evaluation protocol for face, person and vehicle detection & tracking in video analysis and content extraction (VACE-II), University of South Florida (Jan 2006).
- [21] M. Isard and A. Blake: CONDENSATION - conditional density propagation for visual tracking, *Int. J. Computer Vision*, 29, (1998), 5-28.