A Conservative Scene Model Update Policy

Nick Mould and Joseph P. Havlicek
University of Oklahoma
School of Electrical and Computer Engineering
Norman, OK, USA
nick.mould@gmail.com, joebob@ou.edu

***Abstract*—In this paper, we present a new pixel-level scene model for segmenting video into foreground and background structure. The design of the model is largely influenced by several recently reported stochastic background models that have been shown to significantly outperform traditional deterministic techniques. In contrast to existing nonparametric scene models, we propose a learning algorithm that integrates new information into the models by replacing the most significant outlying values with respect to the current sample collections. Outliers are identified using a variable bandwidth kernel density estimation (KDE) procedure. We demonstrate the superiority of our model against a recent state-of-the-art video segmentation system and compare and contrast the theoretical aspects of our model with a wide variety of existing techniques, and well known video segmentation challenges.**

***Index Terms*—video segmentation, scene modeling, background modeling**

I. Introduction

We consider the problem of segmenting video into foreground and background regions using pixel level scene modeling techniques. Recently, the first nondeterministic background model (ViBe) was proposed in [1] and shown to outperform a wide variety of existing algorithms. The ViBe model is unique in that it is the first and only scene model that uses a completely stochastic maintenance algorithm to integrate new information into the system. We implemented ViBe and immediately observed its superiority to several other well known scene modeling techniques, namely, the GMM of Stauffer and Grimson [2], [3], the multidimensional median filter of [4], the temporal low-pass filter of [5] and the KDE technique proposed by Elgammal, Harwood and Davis in [6], [7]. In [1], Barnich demonstrated the effectiveness of the ViBe model against the Zivkovic GMM [8], the Codebook proposed in [9], a pixel level single Gaussian model with adaptive variance, and several other lesser known techniques such as the $\Sigma - \Delta$ model [10], a Bayesian histogramming algorithm [11], an alternative GMM [12], and a simple temporal low-pass filter similar to [5]. In addition, Brutzer [13] independently verified the claims of Barnich by comparison to another collection of well known scene models that included a classical median filter [14], the Stauffer and Grimson GMM [2], [3], the Oliver and Pentland Eigenbackground subtraction method [15], the single Gaussian model proposed in [16], a Bayesian histogram [11], the Codebook of [9], the Zivkovic GMM [8], and a self organizing map (SOM) [17].

The ViBe scene model is a pixel level nonparametric background model that operates in the grayscale or RGB colorspaces and uses kernel density estimation (KDE) to classify pixels in unsegmented video frames. The number of previously observed samples that are used to characterize the distributions of background values at each pixel location is fixed at twenty. The background probabilities of each pixel in an unsegmented frame are estimated by performing KDE using a spherical cutoff kernel [1] with a fixed radius of twenty pixels. If the background probability is less than or equal to 0.1, then the pixel is classified as foreground, otherwise it is classified as background and integrated into the system at the pixel level and possibly at the neighborhood level.

In ViBe, pixels that are classified as background are automatically inserted into the sample collection at the corresponding pixel location. In contrast to existing nonparametric models where the oldest value in the sample collection would be replaced by the new value, ViBe uses a uniformly distributed random variable to determine the index of the sample to be replaced. The authors show that this policy ensures that the expected lifespan of each sample decays exponentially and that the probability of a sample being preserved is independent of time, and therefore that the system is memoryless. We propose a different update policy that replaces the most significant outlier in the sample collection. We argue that the outlying value is both the least important sample in the statistical sense, and the most likely sample to represent a component of the foreground that has been erroneously included in the background model. In Section VI we show that this outlier replacement policy has no negative impact on the performance of the algorithm and that it nearly eliminates the persistent ghost problem described later in this section.

In addition to integrating background pixels into the corresponding pixel level models, the values may be propagated to a single neighboring distribution to promote spatial consistency throughout the scene. In the ViBe system, a uniformly distributed random variable is used to propagate the background sample to a neighboring model with a probability of $1/16$. In the case that the sample is selected for propagation, one of the eight neighboring sample collections is randomly selected using another uniformly distributed random variable. Selection of the sample within the neighboring distribution to be replaced is also performed by a stochastic process where a random variable that follows a uniform distribution is used to determine which sample to replace. The ViBe neighborhood diffusion process is based on an assumption that balances two conflicting premises, namely that the structures of the neighboring distributions are similar enough that information can be randomly swapped without fear of corrupting the sample collections, yet disparate enough that swapping of information improves the diversity of each model in a constructive sense. In the case where neighboring pixel level models lie on a different side of an edge boundary, the assumption that adjacent distributions are similar is clearly incorrect and will lead to unpredictable corruption of the two models through the diffusion algorithm of [1]. Because the neighboring substitution index is chosen at random, the potential for severe damage to the model is greatly increased as important and unimportant values are equally likely to be replaced. Indeed, all of our simulations with the model from [1] revealed unjustifiably high foreground probabilities along the edges of the background structure when examined prior to application of the final segmentation threshold. The proposed outlier replacement strategy reduces the effects associated with random propagation of information to neighboring models because the impact to the model is minimized by replacement of the least significant sample.

We identify the following four scene modeling components and use them to describe the theoretical aspects of our algorithm and to compare and contrast the model with a wide variety of existing techniques.

- **Model representation:** The collection of static and dynamic system parameters combined with data storage elements that represent the model at a single discrete time instance k.
- **Model initialization:** The method by which the elements of the scene model are initialized at time $k = 0$.
- **Frame segmentation:** The procedure used to compare an unsegmented video frame to the current instance of the model to arrive at a segmented video frame.
- **Model maintenance:** The algorithm or update policy used to integrate new information into the existing scene model. The maintenance strategy may or may not make use of the segmented frame, but in general it will make use of the image features observed within the observed unsegmented video frame.

In addition, we provide the reader with a descriptive list of the challenging problems and definitions thereof that have been historically encountered in the field of video segmentation by aggregating the work of [18] and [13] in the following comprehensive collection. From this point forward, we use these terms to analyze both the theoretical aspects of our proposed algorithm as well as the simulation results.

- **Bootstrapping:** In many situations, the scene model must be initialized in the presence of foreground objects, and because a trusted model of the scene does not yet exist, it is impossible to determine the difference between foreground and background objects. In the video segmentation literature, this procedure is known as bootstrapping, although the actual statistical term "bootstrap" is at best only loosely related to this process.
- **Gradual illumination changes:** Reasonable changes in lighting conditions such as those that are naturally occurring and expected in outdoor environments.
- **Sudden illumination changes:** Unexpected variations in lighting conditions that occur frequently in indoor settings, but are generally unpredictable.
- **Dynamic background components:** Swaying tree branches, rippling water, and uninteresting components of the scenery are all common examples of dynamic background components. Unfortunately, the definitions of background and foreground are not completely straightforward, and thus, the term background may refer to any elements of the scenery that are unimportant to the application at hand.
- **Camouflaged foreground components:** Foreground objects that share very similar color and textural appearance with the background, making detection difficult if not impossible.
- **Shadows:** Shadows may be cast by either foreground or background objects and they pose a significant challenge to video segmentation systems because they generally appear different from the known background components and thus they are incorrectly identified as foreground objects.
- **Ghosts/waking person:** When background objects suddenly become a part of the foreground such as in the case of a parked car leaving its space, the region uncovered by the object is, in many cases, incorrectly identified as a foreground object. If the incorrectly classified region is not quickly identified as part of the background in the model update step, then the object may linger for a long period of time and continue to appear as a persistent ghost.
- **Foreground aperture:** The situation in which homogeneously colored or textured regions within a moving foreground object are incorrectly identified as background structure because they do not appear to be in motion.

II. Model Representation

We employ a pixel level nonparametric model to characterize the temporal distributions of background image features according to [1], [6], [7]

$$M(\mathbf{p}) = \{\phi_1, \phi_2, \phi_3, \ldots, \phi_N\}, \tag{1}$$

where M is a nonparametric model of the background scene represented by a collection of N previously observed values in the grayscale intensity feature space and $\mathbf{p} = (x_1, x_2)$ are the horizontal and vertical coordinates of a single pixel. In terms of versatility, nonparametric models are unique in that they are well suited to the representation of multimodal statistical distributions where the number of modes is unknown and likely to change over time.

Historically, nonparametric models have been shown to provide excellent characterizations of highly dynamic background components and gradual variations in lighting conditions [1], [6], [7], [9], [19]–[23]. Naturally occurring changes in lighting conditions have been easily modeled with unimodal techniques. However, it is impossible to model dynamic background components simultaneously undergoing changes in lighting conditions with unimodal statistical models. Thus, nonparametric techniques have been generally accepted as a powerful tool in the modeling of complex outdoor environments [1], [6], [7].

III. Model Initialization

We performed a blind initialization of the model over N frames, by assigning each grayscale value directly according to

$$\begin{aligned} M(\mathbf{p}) &= \{\phi_1, \phi_2, \phi_3, \ldots, \phi_N\} \\ &= \{I_{k-(N-1)}(\mathbf{p}), \ldots, I_k(\mathbf{p})\}, \end{aligned} \tag{2}$$

where I_k represents a single video frame at time k. Because descriptive information about the foreground and background structures is not generally available during the initialization process, and because the presence of moving foreground objects is both likely to occur and unlikely to be detected accurately, we elected to use a naive initialization strategy. With this approach, the effects of a moving object are spread over several spatial locations rather than concentrated at a single location as in the case of the single frame bootstrapping techniques.

In the ViBe model, initialization is performed by single frame bootstrapping and the samples are randomly selected from a 3×3 neighborhood centered about the model location using a uniformly distributed random variable [1]. Unfortunately, this tactic increases the degree to which moving foreground objects corrupt the initial background model, because entire regions within the model will contain only foreground values. When the video processing begins, these moving foreground regions will begin to uncover the true background structure, resulting in both a true foreground detection due to the moving object and a false foreground detection or ghost in the place of the objects original position. In addition, the random selection of values from a neighborhood may cause neighboring values from significantly different image regions to dominate or perturb the initial model of the background scene in an undesirable way. For these reasons, we have adopted a simpler initialization method that avoids the accidental creation of a ghost and delays the neighborhood diffusion process until sufficient models of the foreground and background structure are available for use in the information sharing process.

IV. Frame Segmentation

Segmentation was performed by thresholding the estimated background probabilities of each observed pixel $I_k(\mathbf{p})$ value within the unsegmented frame I_k according to

$$L_k(\mathbf{p}) = \begin{cases} \text{Foreground}, & P(I_k(\mathbf{p})) < T \\ \text{Background}, & \text{Otherwise} \end{cases}, \tag{3}$$

where T is a fixed threshold and $P(I_k(\mathbf{p}))$ is the background probability of a single observed pixel $I_k(\mathbf{p})$ estimated by

$$P(I_k(\mathbf{p})) = \frac{1}{N}\sum_{i=1}^{N} K(I_k(\mathbf{p}), \phi_i^{\mathbf{p}}). \tag{4}$$

In (4), $\phi_i^{\mathbf{p}}$ represents the i'th sample from the background model M at pixel location $\mathbf{p}$, and K is a uniform spherical cutoff kernel of radius R defined by [1]

$$K(a,b) = \begin{cases} 1, & |a-b| \leq R \\ 0, & \text{Otherwise} \end{cases}, \tag{5}$$

where $a, b \in \mathbb{R}$.

Pixel level segmentation techniques produce high resolution binary classification of foreground and background structures within video. In terms of the foreground aperture problem, these rich segmentations make it possible to use post segmentation algorithms to identify foreground details that penetrate the occluding background structures and use them to reconstruct a more accurate estimate of the object shape. Popular pixel level scene models that have featured post segmentation algorithms for dealing with the foreground aperture problem are the GMM of Stauffer and Grimson [2], [3], where foreground detections are combined through a connected components algorithm, and the nonparametric models of Elgammal, Harwood and Davis [6], [7], where foreground regions are refined through a probabilistic analysis of the neighboring pixels. Not surprisingly, the advantage of high resolution segmentations is not completely without a few drawbacks, namely the susceptibility of pixel level algorithms to the foreground aperture problem. To combat the foreground aperture problem, a wide variety of post segmentation procedures have been proposed, such as a region growing operation by back-projection [18], morphological operations combined with a binary support map to strictly define the support of each foreground object [24], [25], and a probabilistic region growing algorithm [6], [7]. In the model that we propose in this paper, we do not perform any post segmentation processing, however. Because information is shared among neighboring models through the model update policy, the effects of foreground aperture and camouflage on the final segmentations are significantly reduced.

V. Model Maintenance

Here, for the first time, we propose a scene model update policy where pixels that have been identified as foreground in the segmentation step are integrated into the existing pixel level models by replacing the most significant outlying samples. We define the outlier in each background model to be the least probable value by estimating the probability of each sample with respect to the entire sample collection using KDE according to

$$\text{Outlier Index} = \underset{i=1,\ldots,N}{\arg\min}\, \frac{1}{N}\sum_{j=1}^{N} K(\phi_i^{\mathbf{p}}, \phi_j^{\mathbf{p}}), \tag{6}$$

where $\phi_i^{\mathbf{p}}$ and $\phi_j^{\mathbf{p}}$ are samples from the model $M(\mathbf{p})$ and K is a spherical cutoff kernel similar to (5). In (6), the radius of the kernel is computed from the data using the method originally presented by Elgammal, Harwood and Davis in [6], where the bandwidth is set to the median absolute deviation measured between all of the possible sample pairs and where pairs composed of identical samples are excluded. We adopt the neighborhood diffusion process from [1] and randomly select a neighboring model using a random variable that follows a uniform distribution. Once a neighboring distribution is selected, the value is integrated into the model using the outlier replacement strategy described in (6).

This update policy achieves excellent results against the ghost problem, because the image features associated with ghosts generally correspond to outliers in the background sample collections. With respect to the overarching problem of false foreground detections, this outlier replacement policy ensures that the neighboring distributions are only minimally transformed by the diffusion procedure, which is of utmost importance in cases where the adjacent model has been poorly chosen.

VI. Results and Conclusions

We demonstrate the effectiveness of our proposed algorithm on a surveillance video provided by the performance evaluation in tracking and surveillance (PETS) workshop [26]. We identified a 200 frame subsequence of the video that contains a moving person that leaves a ghost behind, and a moving vehicle. Ground truth data for each frame was obtained by manually segmenting each frame of the PETS video. Our initial results are shown in Fig. 1, where (a) and (e) are the original raw video frames, (b) and (f) are the ground truth frames, (c) and (g) are the foreground probability images from the ViBe algorithm, and (d) and (e) are the foreground probability images from our proposed algorithm. In Fig. 1 (c) and (g), erroneously high foreground probabilities are observed along all of the stationary edges due to the uniform replacement of values within poorly chosen neighboring models. In our results, these false foreground detections along with the ghost of the original location of the person are nearly eliminated (Fig. 1 (d) and (h)).

Table I shows the results of the two algorithms evaluated in terms of percentage correct classification (PCC) and a new probability correct classification (PrCC) measurement proposed here for the first time. Percentage correct classification (PCC) is computed according to

$$\text{PCC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{7}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. To better identify the differences in the two scene models, we propose the probability of correct classification (PrCC) measurement and use it to evaluate each algorithm prior to application of the final segmentation threshold. The PrCC is computed according to

$$\text{PrCC} = \frac{\text{TP}_{\text{prob}} + \text{TN}_{\text{prob}}}{\text{TP}_{\text{prob}} + \text{TN}_{\text{prob}} + \text{FP}_{\text{prob}} + \text{FN}_{\text{prob}}} \tag{8}$$

where TP_{prob} is the sum of the foreground probabilities at the ground truth foreground pixel locations, TN_{prob} is the sum of the background probabilities at the ground truth background locations, FP_{prob} is the sum of the foreground probabilities at the ground truth background locations and FN_{prob} is the sum of the background probabilities at the ground truth foreground location. Because a principled threshold selection process does not exist for these types of models, it is better to study the accuracy of the model prior to the use of empirically determined thresholds. In both quantitative assessments, the proposed algorithm outperforms ViBe, and in the case of the model based comparison (PrCC), the improvement is significant.

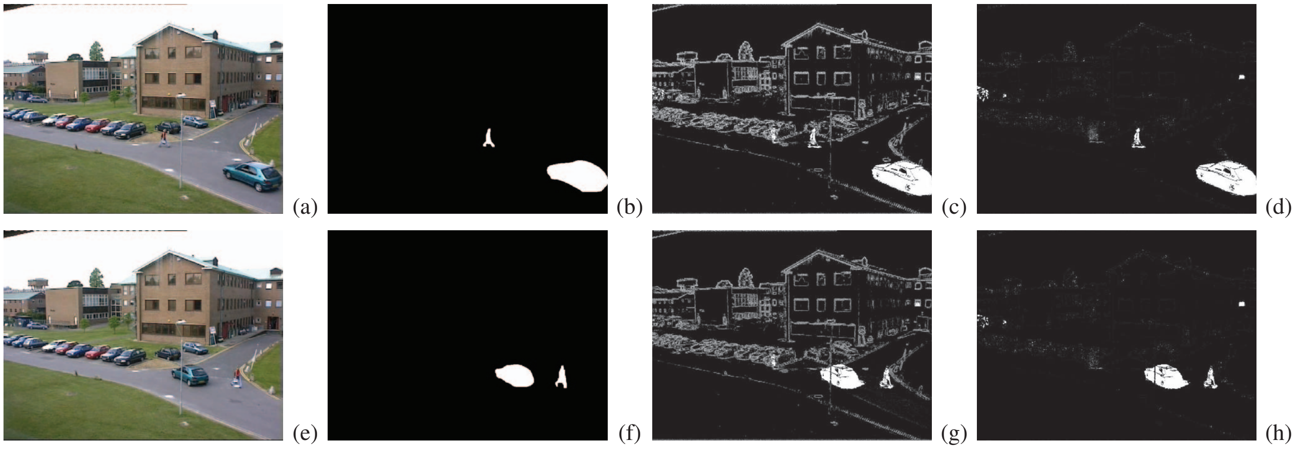

Fig. 1. Selected video frames depicting frames 500 and 600 of the PETS sequence in terms of the original images (a, e), ground truth segmentation images (b, f), ViBe foreground probability images (c, g) and the foreground probability images of the proposed algorithm (d, h).

TABLE I
PERFORMANCE RESULTS OF THE TWO VIDEO SEGMENTATION SYSTEMS.

Model	PCC	PrCC
ViBe	99.3477%	93.5280%
Proposed Algorithm	99.4938%	99.0120%

ACKNOWLEDGMENT

This work was supported in part by the U.S. Army Research Laboratory and the U.S. Army Research Office under grant W911NF-08-1-0293.

REFERENCES

[1] O. Barnich and M. Van Droogenbroeck, “Vibe: A universal background subtraction algortihm for video sequences,” *IEEE Trans. Image Processing*, vol. 20, no. 6, pp. 1709–1724, June 2011.

[2] C. Stauffer and W.E.L. Grimson, “Adaptive mixture models for real-time tracking,” in *Proc. IEEE Int’l. Conf. on Comp. Vision, Pattern Recog.*, Fort Collins, CO, USA, June 23-25 1999, vol. 2.

[3] C. Stauffer and W.E.L. Grimson, “Learning patterns of activity using real-time tracking,” *IEEE Trans. Pattern Anal., Machine Intel.*, vol. 22, no. 8, pp. 747–757, Aug. 2000.

[4] R. Cucchiara, M. Piccardi, and A. Prati, “Detecting moving objects, ghosts, and shadows in video streams,” *IEEE Trans. Pattern Anal., Machine Intel.*, vol. 25, no. 10, pp. 1337–1342, Oct. 2003.

[5] G.W. Donohoe, D.R. Hush, and N. Ahmed, “Change detection for target detection and classification in video sequences,” in *Proc. IEEE Int’l. Conf. Acoustic Speech and Signal Processing*, 1988, pp. 1084–1087.

[6] A.M. Elgammal, D. Harwood, and L. Davis, “Non-parametric model for background subtraction,” in *Proc. European Conf. Computer Vision*, 2000, vol. 1843, pp. 751–767.

[7] A. Elgammal, R. Duraiswami, D. Harwood, and L. Davis, “Background and foreground modeling using nonparametric kernel density estimation for visual surveillance,” *Proc. IEEE*, vol. 90, no. 7, pp. 1151–1163, July 2002.

[8] Z. Zivkovic, “Improved adaptive gaussian mixture model for background subtraction,” in *Proc. IEEE Int’l. Conf. on Pattern Recognition*, 2004, vol. 2, pp. 28–31.

[9] K. Kim, T. Thanarat, H. Chalidabbhognse, D. Harwood, and L. Davis, “Real time foreground-background segmentation using codebook model,” *Real-Time Imaging*, vol. 11, pp. 172–185, 2005.

[10] A. Manzanera, “$\Sigma - \Delta$ background subtraction and the zipf law,” *Progress in Pattern Recognition, Image Analysis and Applications, Springer*, vol. 4756, pp. 42–51, Nov. 2007.

[11] L. Li, W. Huang, I. Gu, and Q. Tian, “Foreground object detection from videos containing complex background,” in *ACM Int’l. Conf. Multimedia*, Berkeley, CA, Nov. 2003, pp. 2–10.

[12] P. KaewTraKulPong and R. Bowden, “An improved adaptive background mixture model for real-time tracking with shadow detection,” in *Proc. European Workshop Advanced Video Based Surveillance Systems*, 2001.

[13] S. Brutzer, B. Hoferlin, and G. Heidemann, “Evaluation of background subtraction techniques for video surveillance,” in *Proc. IEEE Int’l. Conf. on Comp. Vision, Pattern Recog.*, Colorado Springs, CO, June 2011, pp. 1937–1944.

[14] N. McFarlane and C. Schofield, “Segmentation and tracking of piglets in images,” in *Machine Vision and Applications*, 1995, vol. 8(3), pp. 187–193.

[15] N.M. Oliver, B. Rosario, and A.P. Pentland, “A Bayesian computer vision system for modeling human interactions,” *IEEE Trans. Pattern Anal., Machine Intel.*, vol. 22, no. 8, pp. 831–843, Aug. 2000.

[16] S.J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, “Tracking groups of people,” in *Computer Vision and Image Understanding*, Oct 2000, vol. 80(1), pp. 42–56.

[17] L. Maddalena and A. Petrosino, “A self-organizing approach to background subtraction for visual surveillance applications,” *IEEE Trans. Image Processing*, vol. 17, no. 7, pp. 1168–1177, July 2008.

[18] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, “Wallflower: Principles and practice of background maintenance,” in *Proc. IEEE Int’l. Conf. Computer Vision*, Kerkyra, Sep. 20-27 1999, pp. 255–261.

[19] M. Seki, T. Wada, H. Fuliwara, and K. Sumi, “Background subtraction based on cooccurrence of image variations,” in *Proc. IEEE Int’l. Conf. on Comp. Vision, Pattern Recog.*, Madison, WI, USA, June 18-20 2003, vol. 2, pp. 65–72.

[20] A. Mittal and N. Paragios, “Motion-based background subtraction using adaptive kernel density estimation,” in *Proc. IEEE Int’l. Conf. on Comp. Vision, Pattern Recog.*, Washington, DC, USA, June 27 - July 2 2004, vol. 2, pp. 302–309.

[21] Y. Sheikh and M. Shah, “Bayesian modeling of dynamic scenes for object detection,” *IEEE Trans. Pattern Anal., Machine Intel.*, vol. 27, no. 11, pp. 1778–1792, Nov. 2005.

[22] I. Haritaoglu, D. Harwood, and L.S. Davis, “W4: Real-time surveillance of people and their activities,” *IEEE Trans. Pattern Anal., Machine Intel.*, vol. 22, no. 8, pp. 809–830, Aug. 2000.

[23] K. Kim, T.H. Chalidabhongse, D. Harwood, and L. Davis, “Background modeling and subtraction by codebook construction,” in *Proc. IEEE Int’l. Conf. Image Processing*, Oct. 24-27 2004, vol. 5, pp. 3061–3064.

[24] C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, “Pfinder: Real-time tracking of the human body,” *IEEE Trans. Pattern Anal., Machine Intel.*, vol. 19, no. 7, pp. 780–785, July 1997.

[25] Q. Zang and R. Klette, “Robust background subtraction and maintenance,” in *Proc. IEEE Int’l. Conf. on Pattern Recognition*, 2004, vol. 2, pp. 90–93.

[26] Computer Vision Pattern Recognition (CVPR), “Performance evaluation of tracking and surveillance (PETS),” *Website: ftp://ftp.pets.rdg.ac.uk/PETS2001/*, 2001.