Robust automatic target tracking based on a Bayesian ego-motion compensation framework for airborne FLIR imagery

Carlos R. del-Blanco, Fernando Jaureguizar, Narciso García and Luis Salgado

Grupo de Tratamiento de Imágenes, Universidad Politécnica de Madrid, 28040, Madrid, Spain

ABSTRACT

Automatic target tracking in airborne FLIR imagery is currently a challenge due to the camera ego-motion. This phenomenon distorts the spatio-temporal correlation of the video sequence, which dramatically reduces the tracking performance. Several works address this problem using ego-motion compensation strategies. They use a deterministic approach to compensate the camera motion assuming a specific model of geometric transformation. However, in real sequences a specific geometric transformation can not accurately describe the camera ego-motion for the whole sequence, and as consequence of this, the performance of the tracking stage can significantly decrease, even completely fail. The optimum transformation for each pair of consecutive frames depends on the relative depth of the elements that compose the scene, and their degree of texturization. In this work, a novel Particle Filter framework is proposed to efficiently manage several hypothesis of geometric transformations: Euclidean, affine, and projective. Each type of transformation is used to compute candidate locations of the object in the current frame. Then, each candidate is evaluated by the measurement model of the Particle Filter using the appearance information. This approach is able to adapt to different camera ego-motion conditions, and thus to satisfactorily perform the tracking. The proposed strategy has been tested on the AMCOM FLIR dataset, showing a high efficiency in the tracking of different types of targets in real working conditions.

Keywords: Target tracking, Particle Filter, ego-motion, Euclidean transformation, Affine transformation, Projective transformation, geometric transformation distributions, spatiogram, FLIR images, aerial imagery

1. INTRODUCTION

Target tracking in forward looking infrared (FLIR) imagery is an important and challenging subject in military and surveillance applications. In contrast to visual images, FLIR images have low signal-to-noise ratios, target objects low contrasted with the background, and non-repeatability of the target object signature. This fact, along with the competing background clutter, and illumination changes due to weather conditions, make the tracking task extremely difficult. In this context, the prior knowledge about the object dynamics allows to determine the probable locations of the target object, and thus avoiding that the tracking algorithm may be distracted by similar clutter structures. However, for applications based on airborne imagery, the unpredictable camera motion, called ego-motion, induces a global motion in the image that prevents to use the object motion information, which dramatically reduces the tracking performance. This problem is addressed in different manners in the scientific literature, and the different methods can be split into three categories: based on the assumption of low ego-motion, based on the object detection and matching, and based on the ego-motion estimation.

Works that assume a low ego-motion expect that the object maintains its spatio-temporal connectivity along the sequence.[1,2] Since these approaches fail in the case of strong ego-motion, other ones define a search area, centered in the previous object location, where the object is expected to be in the current frame,[3] instead of assuming spatial connectivity. Nevertheless, the probability that the tracking algorithm can be distracted by the clutter background increments with the size of the search area. Instead of making an exhaustive search in a predefined image region, some authors propose to probabilistically model both the camera and the object

E-mail: {cda,fjn,narciso,lsa}@gti.ssr.upm.es
http://www.gti.ssr.upm.es
http://www.gti.ssr.upm.es/∼cda

Automatic Target Recognition XIX, edited by Firooz A. Sadjadi, Abhijit Mahalanobis,
Proc. of SPIE Vol. 7335, 733514 · © 2009 SPIE · CCC code: 0277-786X/09/$18 · doi: 10.1117/12.820203

dynamics by means of linear models using Kalman filters[4] or Particle Filters.[5,6] While these methods could be appropriate for modeling the instability of a camera platform in a stationary situation (for example a camera located on top of a pole that is moving because of wind), the ego-motion arising from an airborne platform is highly non-linear and unpredictable, and therefore the tracking tends to fail in instants related to sudden ego-motion.

Another category of methods propose to detect the target objects, and to perform the matching between them,[7,8] which can cope with arbitrarily large camera motions. However, this approach is totally dependent on the correct object detection, which is a hard task prone to errors.

Works based on the ego-motion estimation offer more versatility for airborne FLIR imagery. They try to compute the camera ego-motion between consecutive frames in order to compensate its side effect, and thus recovering the spatio-temporal correlation of the sequence. The camera ego-motion is assumed to follow a geometric model, usually affine or projective, whose parameters are estimated using an image registration technique.[9] For example, in Refs. 10–14 an area-based image registration technique is used to estimate the parameters of an affine geometric model. However, the presence of independent moving objects can drift the ego-motion estimation, especially when their size is significative in comparison with the size of the regions of highly structured clutter. In this case, a feature-based technique along with robust estimation methods can obtain better results. Nonetheless, due to the aforementioned drawbacks of the FLIR imagery, distinctive features can hardly be detected. Some works[15–17] use edge based features, that are not the most distinctive but can be easily detected, and then, they make use of the robust estimation theory to deal with the large number of outliers in the computation of an affine model.

All ego-motion estimation based approaches have in common that use only one geometric transformation to model the camera motion along the whole sequence. This may not be suitable because the selection of the appropriate geometric model is a tradeoff between the capability of the proposed transformation to model the camera motion, and the accuracy in the estimation of its parameters. While the geometric camera model is projective, the accuracy of the estimation of its eight parameters may become very low due to the poor image quality of the FLIR imagery. If a geometric model with a lower number of parameters (affine or Euclidean models) is selected, the accuracy can increase since there are less degrees of freedom. However, the estimated motion could not properly represent the camera motion, for example when the camera is close to the target object, an Euclidean or affine motion model is unable to capture the skew, pan and tilt of the planar scene. Moreover, the appropriate camera model depends on the depth relief of the objects, the average depth in the scene, and the size of the field of view of the camera,[18] information that usually is not available.

In this work, a novel approach for object tracking under strong camera ego-motion conditions is proposed, which efficiently manages several models of camera ego-motion (Euclidean, affine, and projective transformations) using a Particle Filter framework. For each transformation model, a discrete likelihood distribution is computed, which approximates the space of geometric transformations by a set of weighted samples. These geometric transformation distributions model the camera dynamics, and along with the own object dynamics constitute the system model of a Particle Filter. Both the camera and the object dynamics are used to compute a sampled prior probability distribution function (pdf) of the object location. Based on appearance information, the prior object location pdf is updated according to the measurement model of the Particle Filter. Using a spatiogram to encode the appearance information, the measurement model evaluates the similarity of the spatiogram related to the target object and those ones corresponding to the samples of the prior object location pdf. The resulting posterior object location pdf is used to estimate the object location in each time step. According to this approach, the samples drawn from the geometric transformation that best models the camera ego-motion will have a higher weight. This allows to adapt to different camera ego-motion conditions, and thus to satisfactorily perform the tracking.

The rest of the paper is organized as follows: in Sec. 2 the general Particle Filter framework for object tracking is presented. The system model of the Particle Filter is described in Sec. 3, where the multi-geometric transformation approach for modeling the camera dynamics is explained. Section 4 presents the measurement model of the Particle Filter that uses a spatiogram to encode the appearance information of the object. Experimental results over the AMCOM dataset are exposed in Sec. 5, and final conclusions are presented in Sec. 6.

2. BAYESIAN VISUAL TRACKING FRAMEWORK

The Bayesian tracking aims to estimate the state of an object that changes over time using a sequence of noisy measurements. The state of the object $\mathbf{x}_k$ at time k is a vector that stores the kinematic (position and velocity on the image plane) and geometric (size and shape) information to characterize the object. It is mathematically expressed as

$$\mathbf{x}_k = [\mathbf{l}_k, \dot{\mathbf{l}}_k, \mathbf{s}_k]^\top, \tag{1}$$

where $\mathbf{l}_k = [l_k^x, l_k^y]^\top$ is a vector with the object spatial coordinates, $\dot{\mathbf{l}}_k = [\dot{l}_k^x, \dot{l}_k^y]^\top$ is a vector with the velocity information, and $\mathbf{s}_k = [s_k^M, s_k^m, s_k^\theta]^\top$ is a vector that respectively encodes the mayor axis, the minor axis, and the orientation of an ellipse that encloses the object.

The noisy measurements $\mathbf{z}_k$ is a vector of observations related to the object information contained in $\mathbf{x}_k$. These observations are the intensity data of the image sequence: $\mathbf{z}_k = \mathbf{I}_k$.

The Bayesian approach calculates some degree of belief in the state $\mathbf{x}_k$ at time k, using the prior information about the object, and the set of measurements $\mathbf{z}_{1:k} = \{\mathbf{z}_i, i = 1, ..., k\}$ up to time k. Thus, the tracking problem consists in computing the posterior probability density function (pdf) $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ of the state of the object. It is assumed that the initial pdf $p(\mathbf{x}_0|\mathbf{z}_0) \equiv p(\mathbf{x}_0)$, called the prior, is known. In the present work, $p(\mathbf{x}_0)$ is initialized as a Kronecker's delta function $\delta(\mathbf{x}_0)$ using ground truth information (since it is available with the used test sequences AMCOM). In a general case, $p(\mathbf{x}_0)$ could be initialized as a Gaussian function using the information given by an object detector algorithm, as in Refs. 1, 2, 13–17.

In order to be computationally efficient, $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ is recursively calculated in two stages: prediction and update. The prediction stage involves to obtain the prior pdf $p(\mathbf{x}_k|\mathbf{z}_{1:k-1})$ of the state at time k using the posterior pdf $p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})$ at the previous time step via the Chapman-Kolmogorov equation[19]

$$p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1}) d\mathbf{x}_{k-1}, \tag{2}$$

where $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ is the state transition probability, defined by the equation

$$\mathbf{x}_k = \mathbf{f}_k(\mathbf{x}_{k-1}, \mathbf{v}_{k-1}), \tag{3}$$

called the system model, where $\mathbf{f}_k$ is a stochastic process that models the camera and object dynamics, and $\mathbf{v}_{k-1}$ is an independent stochastic process of noise that models the unknown disturbances in the state prediction. The system model is described in detail in Sec. 3.

The predicted pdf $p(\mathbf{x}_k|\mathbf{z}_{1:k-1})$ appears usually translated, deformed, and spread by the process noise $\mathbf{v}_{k-1}$. The update stage refines $p(\mathbf{x}_k|\mathbf{z}_{1:k-1})$ using the new measurement $\mathbf{z}_k$ (measurements are assumed to be available at discrete times) through the Bayes' rule

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) = \frac{p(\mathbf{z}_k|\mathbf{x}_k) p(\mathbf{x}_k|\mathbf{z}_{1:k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})}. \tag{4}$$

The likelihood function $p(\mathbf{z}_k|\mathbf{x}_k)$ relates the noisy measurements to the state of the object, and it is defined by the equation

$$\mathbf{z}_k = \mathbf{h}_k(\mathbf{x}_k, \mathbf{n}_k), \tag{5}$$

called the measurement model, where $\mathbf{h}_k$ is a stochastic process that evaluates in what degree the measurement supports the prediction, and $\mathbf{n}_k$ is an independent stochastic process related to the measurement noise. The measurement model is described in detail in Sec. 4.

The denominator of Eq. (4) is a normalizing constant given by

$$p(\mathbf{z}_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{z}_k|\mathbf{x}_k) p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) d\mathbf{x}_k. \tag{6}$$

The recursive propagation of the posterior density, accomplished by Eqs. (2) and (4), is the optimal solution, but in practice it can not be determined analytically. In this situation, suboptimal methods can be used to approximate the optimal Bayesian solution. Particle Filtering is one of the most popular suboptimal methods, since it is able to deal with continuos state spaces and nonlinear/non-Gaussian processes, in contrast to other suboptimal methods, such as Extended Kalman Filters, Unscented Kalman Filters and Hidden Markov Models, that impose some operational restrictions.

Particle Filtering is a Monte Carlo method for simulating recursive Bayesian filters. In each time step, the posterior density function is approximated by a set of N_s weighted random samples[19]

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) \approx \frac{1}{S_w}\sum_{i=1}^{N_S} w_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i), \tag{7}$$

where the function $\delta(x)$ is the Kronecker's delta, $\{w_k^i, i = 0, ..., N_S\}$ is the set of weights associated to the samples, and $S_w = \sum_{i=1}^{N_S} w_k^i$ is a normalization weight factor. As the number of samples becomes very large, this approximation becomes equivalent to the true posterior pdf, and thus the Particle Filter approaches the optimal Bayesian estimate.

Both samples $\mathbf{x}_k^i$ and weights w_k^i are computed using the principle of the importance sampling,[19,20] which is a simulation technique that aims to reduce the variance of the estimation given by Eq. (7). This is accomplished selecting an appropriate set of samples $\{x_k^i, i = 0, ..., N_S\}$ that are drawn from an alternative distribution function $q(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_k)$, called the importance density. The optimal $q(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_k)$ should be proportional to $p(\mathbf{x}_k|\mathbf{z}_{1:k})$, and have the same support (the support of a function is the set of points where the function is not zero), since in that case the variance is zero. However, this is only a theoretic solution since it would imply the knowledge of $p(\mathbf{x}_k|\mathbf{z}_{1:k})$. A practical and widely adopted solution is to use $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ as the importance density, which is an acceptable approximation provided that it is not much wider than the likelihood $p(\mathbf{z}_k|\mathbf{x}_k)$, and the main modes of $q(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_k)$ do not lie in the tails of $p(\mathbf{z}_k|\mathbf{x}_k)$.

The weights w_k^i related to each sample $\mathbf{x}_k^i$ are recursively computed by[19]

$$w_k^i = w_{k-1}^i \frac{p(\mathbf{z}_k|\mathbf{x}_k^i)p(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i)}{q(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i, \mathbf{z}_k)}. \tag{8}$$

Since $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ embodies all available statistical information, an optimal estimate of the state of the object $\widehat{\mathbf{x}}_k$ may be computed. Assuming that the shape component $\mathbf{s}_k$ of the state of the object varies slowly, which is justified by the rigid nature of target objects, and taking into account the possible multi-modality of $p(\mathbf{x}_k|\mathbf{z}_{1:k})$, an efficient estimator can be derived using the weighted kernel density estimation theory. This is expressed mathematically as

$$\widehat{\mathbf{x}}_k = \max\left(\frac{1}{N_s}\sum_{i=1}^{N_s} p(\mathbf{x}_k^i|\mathbf{z}_{1:k}) K(\mathbf{x}_k - \mathbf{x}_k^i)\right), \tag{9}$$

where $K(\mathbf{x})$ is a multidimensional Gaussian kernel of mean the zero vector, and covariance matrix Σ_n. As it will be seen in Sec. 3, Σ_n is the same as the covariance matrix of the noise stochastic process $\mathbf{v}_{k-1}$. The computed estimation $\widehat{\mathbf{x}}_k$ is similar to the minimum mean square error (MMSE) estimator, but restricting the computation to the set of samples belonging to the principal mode of $p(\mathbf{x}_k|\mathbf{z}_{1:k})$.

The performance of the proposed Bayesian tracking algorithm depends on the appropriate design of the system and measurement models, which are respectively described in Sec. 3 and 4.

3. CAMERA AND OBJECT DYNAMICS

The system model of the Particle Filter framework uses the available prior information to describe the temporal evolution of the state of the object, which depends not only on the own object dynamics, but also on the camera dynamics. This dual dependency arises from the fact that the camera motion induces a global motion in the sequence that shifts the expected object location. According to this, the evolution of the state of the object can be expressed as

$$\mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1}, \mathbf{v}_{k-1}) = \mathbf{f}_{cam}(\mathbf{x}_k^{obj}, \mathbf{g}_{k-1}) \tag{10}$$

where $\mathbf{x}_k^{obj} = \mathbf{f}_{obj}(\mathbf{x}_{k-1}, \mathbf{v}_{k-1})$ is a linear approximation of the object dynamics given by

$$\mathbf{f}_{obj}(\mathbf{x}_{k-1}, \mathbf{v}_{k-1}) = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{v}_{k-1}, \tag{11}$$

where the matrix $\mathbf{A}$ is defined, according to a model of constant velocity and shape, as

$$\mathbf{A} = \left[\begin{array}{c|c|c} \mathbf{I}_2 & \mathbf{I}_2 & \mathbf{0}_{2\times 3} \\ \hline \mathbf{0}_{5\times 2} & \multicolumn{2}{c}{\mathbf{I}_5} \end{array}\right]. \tag{12}$$

The matrices $\mathbf{I}_2$ and $\mathbf{I}_5$ are respectively identity matrices of size 2×2 and 5×5, and $\mathbf{0}_{2\times 3}$ and $\mathbf{0}_{5\times 2}$ are respectively zero matrices of size 2×3, and 5×2.

The i.i.d. noise Gaussian process $\mathbf{v}_{k-1}$ models the unknown disturbances in the linear state prediction, so that the object dynamics can deal with slight variations in velocity and size.

The function $\mathbf{f}_{cam}(\mathbf{x}_k^{obj}, \mathbf{g}_{k-1})$ models the camera dynamics by means of a 2D geometric transformation, which can be of type Euclidean, affine or projective. This geometric transformation is obtained from the stochastic process $\mathbf{g}_{k-1}$, which is described in Subsec. 3.1. Then, each geometric transformation sample $\mathbf{g}_{k-1}^i$, drawn from $\mathbf{g}_{k-1}$, is used to warp the ellipse defined by $\mathbf{x}_k^{obj}$. The resulting $\mathbf{x}_k$ contains the ellipse parameters that predicts the object location in the current frame. Notice that for affine and projective transformations, the warping can produce a non elliptical shape region. In this case, the resulting warped region would be approximated by an ellipse.

3.1 Geometric transformation likelihoods

The stochastic process $\mathbf{g}_{k-1}$ randomly draws a geometric transformation from one of the likelihoods $p(\mathbf{g}_{k-1}^{eu}|I_{k-1:k})$, $p(\mathbf{g}_{k-1}^{af}|I_{k-1:k})$, and $p(\mathbf{g}_{k-1}^{pr}|I_{k-1:k})$, which are respectively the Euclidean, affine and projective likelihoods. These are conditioned to $I_{k-1:k}$, the set of frames acquired between the time steps $k-1$ and k, i.e. between the frames for which the camera motion is estimated. The Euclidean likelihood is approximated by a set of weighted Euclidean transformations $\mathbf{g}_{k-1}^{eu,(i)}$ that represent the most probable camera motions between the instants $k-1$ and k, assuming an Euclidean transformation model. Mathematically, it is expressed as:

$$p(\mathbf{g}_{k-1}^{eu}|I_{k-1:k}) \approx \frac{1}{N_{eu}} \sum_{i=1}^{N_{eu}} w_k^{eu,(i)} \delta(\mathbf{g}_{k-1}^{eu} - \mathbf{g}_{k-1}^{eu,(i)}), \tag{13}$$

where $w^{eu,(i)}, i = 1, ..., N_{eu}$ is the set of weights relative to samples.

The set of Euclidean transformations $\{\mathbf{g}_{k-1}^{eu,(i)}, i = 1, ..., N_{eu}\}$ are obtained from a feature-based correspondence process. Features $\mathbf{u}_{k-1}^j$ are randomly selected in the image I_{k-1} among the regions of higher gradient, which are obtained using a Canny edge detector algorithm. Despite that this set of features can be affected in some degree by the so called aperture problem,[21] it is the most reliable set for the correspondence task, since techniques that use the curvature information (Harris, SIFT) produce poor results in FLIR images (there are almost no detected points because of their low signal to noise ratio). Each feature is characterized by a descriptor vector $\mathbf{d}(\mathbf{u}_{k-1}^j)$ that contains the phase responses of a Gabor filter bank,[22] which is tuned to different frequency scales and orientations. Then, a similarity likelihood map L_M is computed for each feature and the set of edge features

obtained from I_k, also computed by means of the Canny algorithm. The function used to obtain the similarity likelihood measure between the features $\mathbf{u}_{k-1}^j$ and $\mathbf{u}_k^h$ is

$$L_M(\mathbf{u}_{k-1}^j, \mathbf{u}_k^h) = \frac{1}{\sqrt{2\pi}\sigma_d} e^{-\frac{||\mathbf{d}(\mathbf{u}_{k-1}^j), \mathbf{d}(\mathbf{u}_k^h)||^2}{2\sigma_d^2}}, \tag{14}$$

where $||a, b||$ is the Euclidean distance, and σ_d is the expected deviation of the feature descriptor vectors in consecutive images.

Since the best estimated correspondence for each feature does not guarantee the true one, several putative correspondences are associated to each feature, those that have the highest values according to its L_M, and are local maxima. Using this framework of multiple correspondences per feature, Euclidean transformation samples $\mathbf{g}_{k-1}^{eu,(i)}$ can be obtained using the RANSAC algorithm. For this purpose, a random selection of N_{cor} correspondences is made, taking into account that the features involved in correspondences must be different. N_{cor} is chosen to be the minimum required to compute the desired geometric transformation, that in the case of an Euclidean transformation is $N_{cor} = 2$ (3 and 4 for affine and projective transformations, respectively). The weight associated to each Euclidean transformation sample $\mathbf{g}_{k-1}^{eu,(i)}$ is computed as

$$w^{eu,(i)} = \sum_{j=1}^{N_f} L_M(\mathbf{u}_{k-1}^j, \mathbf{g}_{k-1}^{eu,(i)} \mathbf{u}_{k-1}^j), \tag{15}$$

where N_f is the number of features in I_{k-1}. The weight will be higher for those Euclidean transformations that align or nearly align the frames $I_{k-1:k}$, since in that case the 2D points correspond to the same 3D point in the scene.

The affine and projective transformation likelihoods, respectively $\mathbf{g}_{k-1}^{af}$ and $\mathbf{g}_{k-1}^{pr}$, are computed in a similar way to the Euclidean case.

Figure 1 intuitively shows the resulting geometric transformation likelihoods using the pair of FLIR images corresponding to Fig. 2. The dashed black lines that appear in these FLIR images have been drawn as a visual aid to clearly observe the camera motion. Figure 1 is composed by three different scattered plots, corresponding from left to right to the affine, Euclidean and projective transformation likelihoods. Circles represent the result of warping a reference 2D point $\mathbf{p}(x, y) = (1, 1)$ by the transformation samples of each transformation likelihood, and the cross marks the reference point $\mathbf{p}(x, y)$. In an ideal case, without independent moving objects and erroneous estimations of transformations, it would be expected to find high concentration of circles (i.e. a mode) around a point, which represents the warping of the transformation that best models the camera motion. However, in a real case, several distorted and spread modes could appear. In Fig. 1, a mode can be distinguished over the point $\mathbf{q}(x, y) = (1, 8)$ for the Euclidean case. The same mode can also be observed for the affine likelihood, but it is more spread due to two reasons. The first one is that the uncertainty in the estimation grows with the number of estimated parameter, and the second one is derived from the RANSAC method used in the transformation likelihood estimation, since it is more complicated to randomly select three correspondences without outliers (affine case) than two ones (Euclidean case). For the same reasons, the circle distribution in the projective case is still more distorted and more spread. Notice that only the transformation samples have been taken into account, but not their weights. Combining both informations, the main mode would be more distinguishable. Figure 3 shows the weight distribution for each transformation model, which approximately follow a Gaussian distribution.

Figure 4 illustrates the fact that the most appropriate transformation model for representing the camera motion varies along the time. The first row shows the absolute difference between the frame I_{k_1}, and the images $I_{k_1-1}^{eu}$, $I_{k_1-1}^{aff}$, and $I_{k_1-1}^{proj}$, which are obtained by warping the frame I_{k_1-1} with the transformations that have the highest weights in the Euclidean, affine, and projective likelihoods, respectively. Note that darker values correspond to lower values. The second and the third row show the same information, but for different consecutive time steps (k_2 and k_3). In the first row, it can be visually observed that the Euclidean transformation is the best one in representing the camera motion, which is quantitatively confirmed by its superior Peak Signal to Noise

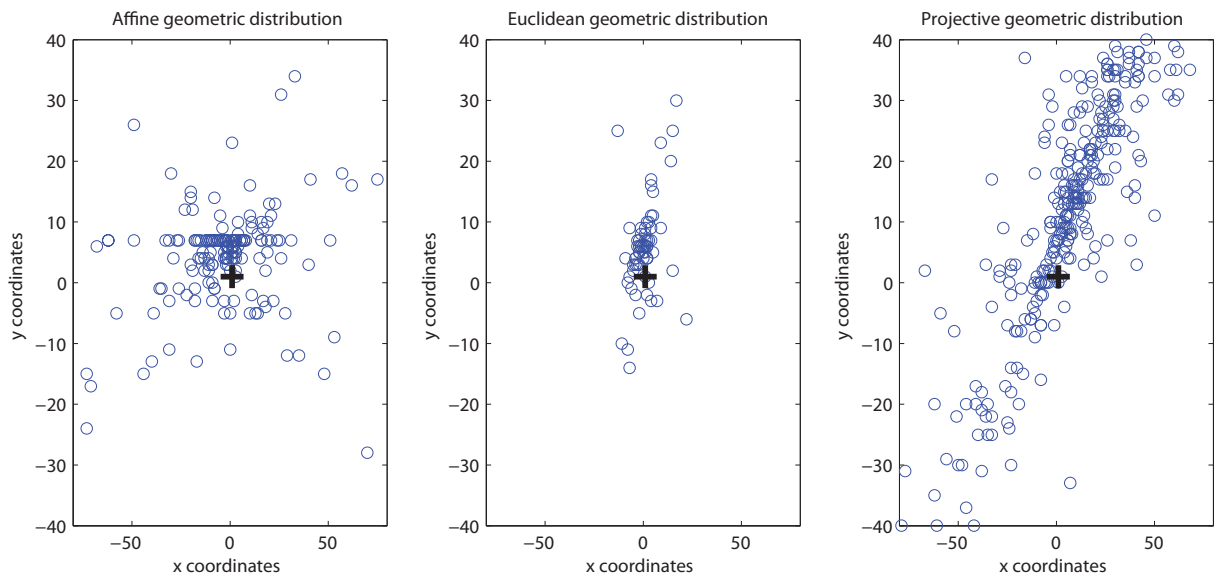


Figure 1. Experiment to intuitively show the affine, Euclidean and projective transformation likelihoods corresponding to the images of Fig. 2. Circles represent the warpings of a reference point, marked by a cross, using the transformation samples that represent the transformation likelihoods.

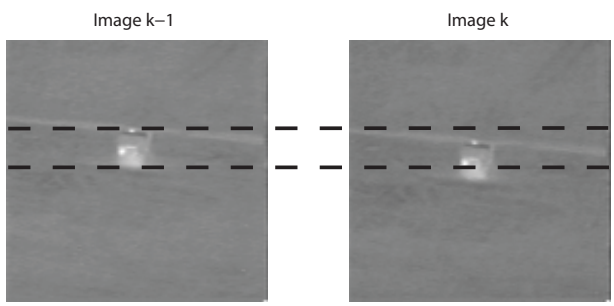


Figure 2. A pair of consecutive FLIR images. The dashed black lines have been drawn as a visual aid to clearly observe the camera motion.

Ratio (PSNR) measure (it appears under each difference image). For the second and third row, the projective and affine transformations are respectively the most appropriate to represent the camera motion.

4. OBJECT MODEL AND SIMILARITY LIKELIHOOD

The target object is modeled by means of its appearance using a spatiogram.[23] Spatiograms are histograms that have been augmented with spatial information to capture a richer description of the object. Considering an intensity image, the spatiogram of a region is a vector whose components are defined by $h(b) = [n_b, \mu_b, \Sigma_b], b = 1, ..., N_b$, where n_b is the number of pixels contributing to the bth bin (similar to a histogram), μ_b and Σ_b are respectively the mean vector and covariance matrix of the spatial coordinates of the pixels belonging to the bth bin. As a result, the object is not only characterized by its intensity distribution, but also by some basic shape information. Notice that the mean and covariance of each bin define an ellipse that represents the spatial distribution of the pixels contributing to that bin.

The distance measure between two spatiograms h and h' is defined by means of the Bhattacharyya distance $d_S(h, h') = \sqrt{1 - \rho_S(h, h')}$, where $\rho_S(h, h')$ is a weighted version of the Bhattacharyya coefficient given by

$$\rho_S(h, h') = \Sigma_{b=1}^{N_b} w_b \rho_B(n_b, n'_b), \tag{16}$$

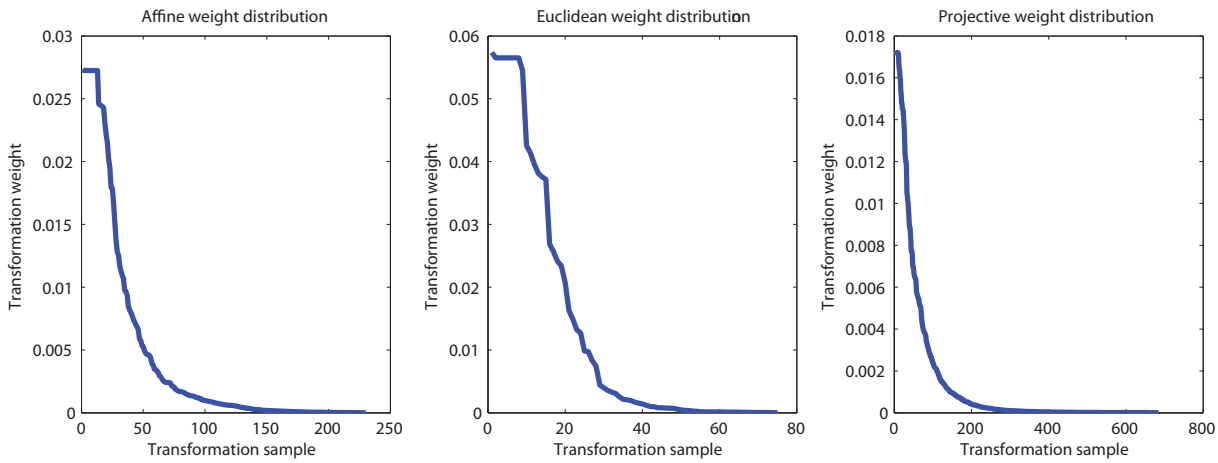


Figure 3. Weight distributions of the samples of the affine, Euclidean and projective transformation likelihoods.

where $\rho_B(n_b, n'_b) = \sqrt{n_b n'_b}$ is the the Bhattacharyya coefficient, and the weights are defined by

$$w_b = N(\mu_b, \mu'_b, \Sigma'_b) N(\mu'_b, \mu_b, \Sigma_b), \tag{17}$$

where $N(n, \mu, \Sigma)$ is a multivariate Gaussian function evaluated at n.

The measurement model evaluates the likelihood that a candidate region defined by $\mathbf{x}_k^i$, and characterized by its spatiogram h_c corresponds to the tracked object, whose appearance is modeled by the spatiogram h_o. This is mathematically expressed as

$$p(\mathbf{z}_k|\mathbf{x}_k^i) = \frac{1}{\sqrt{2\pi}\sigma_S} \exp - \left(\frac{1 - \rho_S(h_c, h_o)}{2\sigma_S^2} \right), \tag{18}$$

where σ_S is the expected variation of the Bhattacharyya distance due to the temporal evolution of the object appearance.

5. RESULTS

The proposed object tracking algorithm has been tested using the AMCOM dataset. This consists of 40 infrared sequences acquired from a camera mounted on an airborne platform. A variety of moving and stationary terrestrial targets can be found in two different wavelengths: midwave (3μm - 5μm) and longwave (8μm - 12μm). In general, the tracking task is quite challenging in this dataset due to the strong camera ego motion, the magnification and pose variations of the target signatures, and the own characteristics of the FLIR imagery described in Sec. 1.

Figures 5, 6, and 7 show some tracking results for the sequence 'rng14_15' of the AMCOM dataset. Figure 5 shows the tracking of a stationary object in some representative frames of the sequence. The target object is enclosed by an ellipse defined by $\widehat{\mathbf{x}}_k$, i.e. the object state estimation in that time step. The proposed tracking algorithm satisfactorily tracks the target object in the whole sequence, in spite of the strong ego-motion. Figure 6 illustrates how the combination of the camera and the object dynamics can efficiently handle the ego motion. It shows two frames and two set of crosses related to each frame, which are the ellipse centers of each predicted object location according to the system model. The set of crosses in frame (a) have been computed considering only the object dynamics. This results in a poor object location estimation, since due to the camera motion the majority of the crosses have been moved outside the object region. Taking into account the camera dynamics, the side effect of the camera motion can be removed, which improves the accuracy of the predicted object location, as shown in frame (b), where both the object and the camera dynamics have been considered. Figure 7 shows a graph with the spatial distance between the estimated object location (represented by the ellipse center) and the

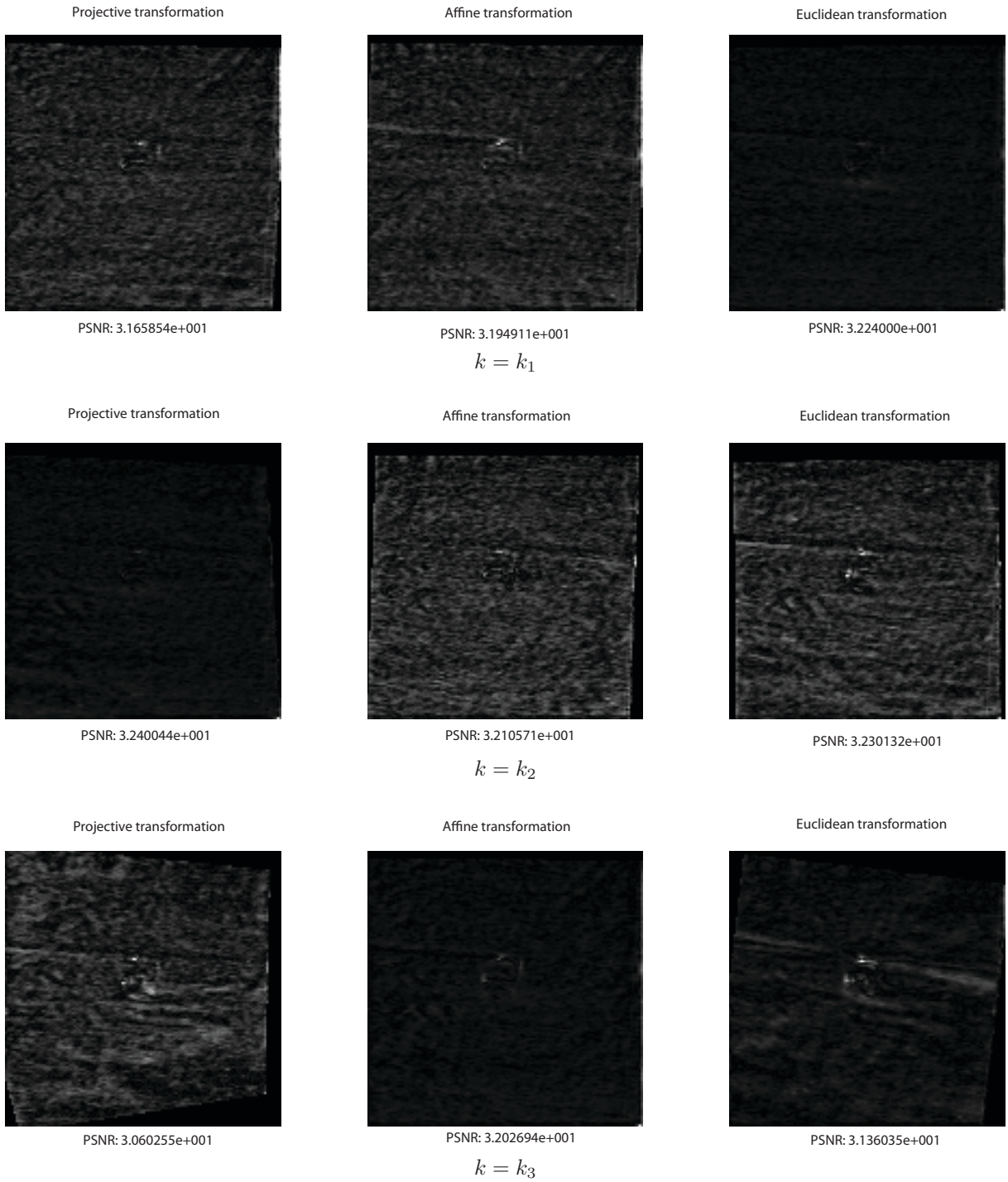


Figure 4. Difference images between a frame and the previous frame warped by a geometric transformation. In each row, a different pair of consecutive frames have been used to compute the difference images, while in each column a different model of transformation has been applied to warp the frame. The difference images along with the PSNR measures reflect that the geometric model that best represents the camera motion varies along the time.

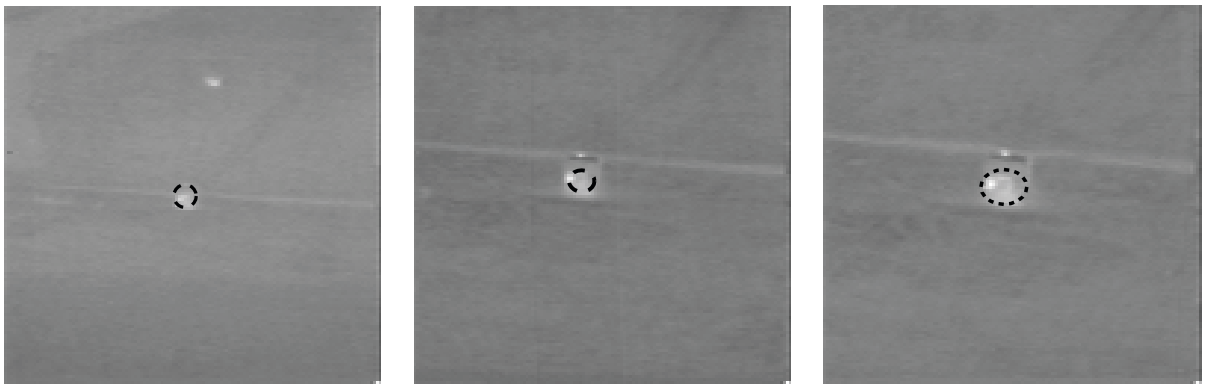
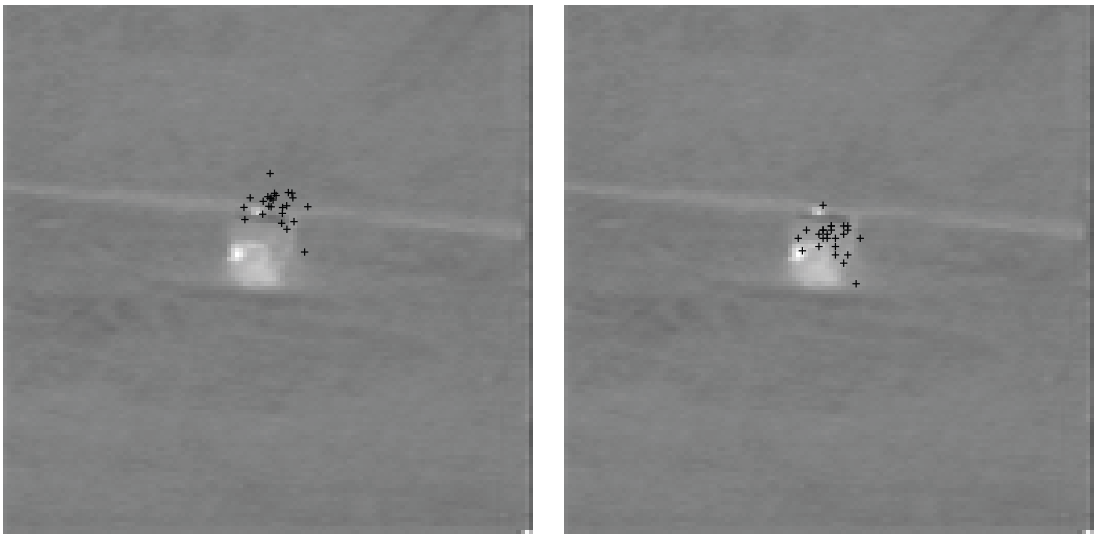


Figure 5. Some tracking results using the sequence 'rng14_15' of the AMCOM dataset. The tracked stationary object is marked by a dashed ellipse.



(a) (b)

Figure 6. (a) Object location prediction, marked with crosses, considering only the object dynamics, (b) Object location prediction considering both the object and the camera dynamics, frame.

available ground truth. The tracking is quite accurate along the time, except for the last part of the sequence, where the increasing object size is not well estimated. This can also be visually appreciated in Fig. 5.

Finally, the performance of the tracking algorithm has been also tested with moving objects. Figure 8 shows a moving object correctly tracked in different time steps of the sequence 'rng16_07' of the AMCOM dataset.

6. CONCLUSIONS

A new approach for object tracking under strong camera ego-motion conditions has been presented. The key idea is to estimate a set of camera motion candidates using several geometric transformations: Euclidean, affine, and projective. These candidates are combined with the prior information about the object motion to yield a robust model of the object dynamics. Then, a Particle Filter framework is used to handle the multiple hypotheses derived from the object dynamics. Thus, the Particle Filter can efficiently predict the object location in conditions of strong ego-motion, where the motion component due to the ego-motion is quite larger than the component due to the moving object. These hypotheses are weighted using the appearance information of the target object. As a result, the hypothesis related to the transformation model that best describes the camera

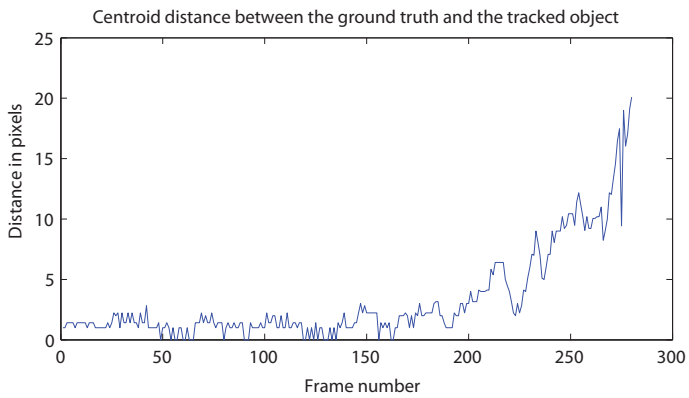


Figure 7. Graphic showing the spatial distance between the estimated object location given by the tracking algorithm, and the available ground truth.

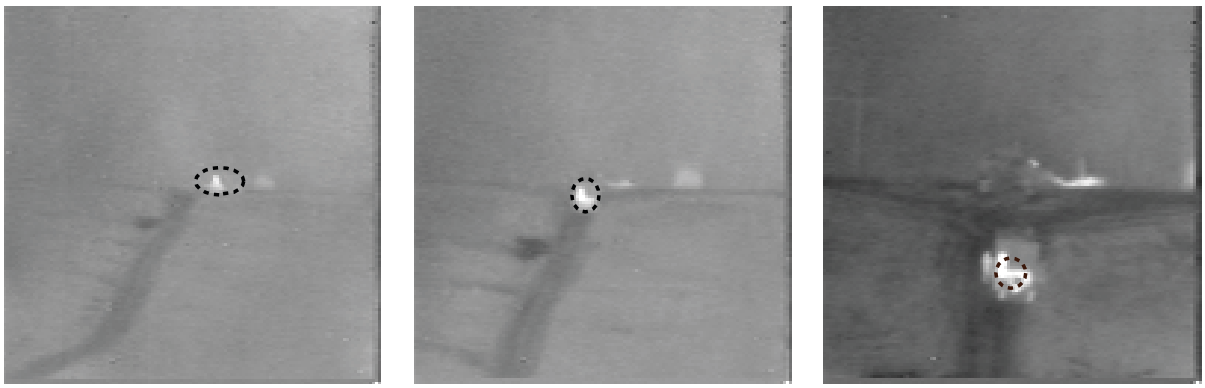


Figure 8. Some tracking results using the sequence 'rng16_07' of the AMCOM dataset. The tracked moving object is marked by a dashed ellipse.

motion will have a higher weight. This allows to efficiently adapt to different camera ego-motion conditions, and thus to satisfactorily perform the tracking. The results obtained using the AMCOM FLIR dataset demonstrate the high performance of the presented tracking strategy in real working conditions, and with different types of targets.

ACKNOWLEDGMENTS

This work has been partially supported by the Comunidad de Madrid under project S-0505/TIC-0223 (Pro-Multidis), and by the Ministerio de Ciencia e Innovación of the Spanish Government under project TEC2007-67764 (SmartVision).

REFERENCES

[1] Braga-Neto, U., Choudhary, M., and Goutsias, J., "Automatic target detection and tracking in forward-looking infrared image sequences using morphological connected operators," *SPIE J. Electronic Imaging* **13**(4), 802–813 (2004).

[2] Xin, H. and Shuo, T., "Target detection and tracking in forward-looking infrared image sequences using multiscale morphological filters," *IEEE Proc. ISPA* , 25–28 (2007).

[3] Bal, A. and Alam, M., "Automatic target tracking in flir image sequences using intensity variation function and template modeling," *IEEE Trans. on Instrumentation and Measurement* **54**(5), 1846–1852 (2005).

[4] Yang, W., Li, J., Shi, D., and Hu, S., "Mean shift based target tracking in flir imagery via adaptive prediction of initial searching points," *IEEE Proc. IITA* **1**, 852–855 (2008).

[5] Mould, N., Nguyen, C., and Havlicek, J., "Infrared target tracking with am-fm consistency checks," *IEEE Proc. SSIAI* , 5–8 (2008).

[6] Venkataraman, V., Fan, G., and Fan, X., "Target tracking with online feature selection in flir imagery," *IEEE Proc. CVPR* , 1–8 (2007).

[7] Shekarforoush, H. and Chellappa, R., "A multi-fractal formalism for stabilization, object detection and tracking in flir sequences," *IEEE Proc. ICIP* **3**, 78–81 (2000).

[8] Alam, M. S. and Bal, A., "Improved multiple target tracking via global motion compensation and optoelectronic correlation," *IEEE Trans. on Industrial Electronics* **54**(1), 522–529 (2007).

[9] Zitova, B. and Flusser, J., "Image registration methods: a survey," *J. Image and Vision Computing* **21**(11), 977–1000 (2003).

[10] Irani, M. and Anandan, P., "Video indexing based on mosaic representations," *IEEE Proc.* **86**(5), 905–921 (1998).

[11] Yilmaz, A., Shafique, K., and Shah, M., "Target tracking in airborne forward looking infrared imagery," *J. Image and Vision Computing* **21**(7), 623–635 (2003).

[12] Yilmaz, A., Shafique, K., Lobo, N., Li, X., Olson, T., and Shah, M. A., "Target-tracking in flir imagery using mean-shift and global motion compensation," *IEEE Proc. CVBVS* , 54–58 (2001).

[13] Strehl, A. and Aggarwal, J., "Detecting moving objects in airborne forward looking infra-red sequences," *IEEE Proc. CVBVS* , 3–12 (1999).

[14] Strehl, A. and Aggarwal, J. K., "Modeep: a motion-based object detection and pose estimation method for airborne flir sequences," *J. Machine Vision and Applications* **11**(6), 267–276 (2000).

[15] del Blanco, C. R., Jaureguizar, F., Salgado, L., and García, N., "Aerial moving target detection based on motion vector field analysis," *Proc ACIVS* **4678**, 990–1001 (2007).

[16] del Blanco, C. R., Jaureguizar, F., Salgado, L., and Garcia, N., "Target detection through robust motion segmentation and tracking restrictions in aerial flir images," *IEEE Proc. ICIP* **5**, 445–448 (2007).

[17] del Blanco, C. R., Jaureguizar, F., Salgado, L., and Garcia, N., "Automatic aerial target detection and tracking system in airborne flir images based on efficient target trajectory filtering," *SPIE Proc. Automatic Target Recognition XVII* **6566**, 656604–1–656604–12 (2007).

[18] R.I., H. and Zisserman, A., [*Multiple View Geometry in Computer Vision*], Cambridge University Press, 2nd ed. (2004).

[19] Arulampalam, S., Maskell, S., and Gordon, N., "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Trans. on Signal Processing* **50**, 174–188 (2002).

[20] Smith, P., Shafi, M., and Gao, H., "Quick simulation: a review of importance sampling techniques in communications systems," *IEEE Journal on Selected Areas in Communications* **15**(4), 597–613 (1997).

[21] Domke, J. and Aloimonos, Y., "A probabilistic notion of correspondence and the epipolar constraint," *Proc. 3DPVT* , 41–48 (2006).

[22] Kamarainen, J. K., Kyrki, V., and Kalviainen, H., "Invariance properties of gabor filter-based features-overview and applications," *IEEE Trans. on Image Processing* **15**(5), 1088–1099 (2006).

[23] Birchfield, S. and Rangarajan, S., "Spatial histograms for region-based tracking," *ETRI Journal* **29**(5), 697–699 (2007).