Object Tracking from Unstabilized Platforms by Particle Filtering with Embedded Camera Ego Motion

Carlos R. del-Blanco, Narciso García, Luis Salgado and Fernando Jaureguizar
Grupo de Tratamiento de Imágenes, Universidad Politécnica de Madrid, 28040, Madrid, Spain
{cda,narciso,lsa,fjn}@gti.ssr.upm.es
http://www.gti.ssr.upm.es http://www.gti.ssr.upm.es/~cda

Abstract

Visual tracking with moving cameras is a challenging task. The global motion induced by the moving camera moves the target object outside the expected search area, according to the object dynamics. The typical approach is to use a registration algorithm to compensate the camera motion. However, in situations involving several moving objects, and backgrounds highly affected by the aperture problem, image registration quality may be very low, decreasing dramatically the performance of the tracking. In this work, a novel approach is proposed to successfully tackle the tracking with moving cameras in complex situations, which involve several independent moving objects. The key idea is to compute several hypotheses for the camera motion, instead of estimating deterministically only one. These hypotheses are combined with the object dynamics in a Particle Filter framework to predict the most probable object locations. Then, each hypothetical object location is evaluated by the measurement model using a spatiogram, which is a region descriptor based on color and spatial distributions. Experimental results show that the proposed strategy allows to accurately track an object in complex situations affected by strong ego motion.

1. Introduction

Visual tracking is a fundamental task in many computer vision applications such as surveillance, autonomous vehicle navigation, robotics, medical imaging, human computer interaction, etc. The aim of the tracking is to localize a previously detected object in each frame of a video sequence. This is essentially accomplished through a correspondence process, consisting in finding the image region that matches closer with the target object, or more specifically with an object model that encodes the main features of the target object. This correspondence process is usually restricted to a subset of image regions, called search area, where it is expected to find the target object according to its dynamics. This allows not only to reduce the computational burden, but also to discard possible false matches, i.e. it simplifies the possible multimodal correspondence process to the unimodal case.

However, in the case of a moving camera (for example mounted on aerial, maritime or terrestrial platform), the tracking may fail since the whole image undergoes a global motion that can move the object outside the search area. To overcome this problem, the typical approach is to use a registration algorithm [10], which estimates the camera motion between consecutive images to compensate it. As a result, it is obtained a sequence of aligned images where the side effect of the camera motion, or ego motion, has been eliminated. The quality of the image registration can decrease dramatically in situations with several independent moving objects, and with backgrounds highly affected by the aperture problem [4]. Despite the robust statistical methods used by some works [9, 1, 8] to address these challenging problems, the enough quality in the image registration to satisfactorily perform the tracking can not be ensured.

Here, a novel approach is proposed to successfully tackle the tracking problem with moving cameras in highly complex situations where other algorithms are prone to fail. The key idea is to compute several highly probable hypotheses of camera motions, instead of trying to deterministically estimate the best one. These hypotheses represent a discrete approximation of the underlying probability distribution function (pdf) of the camera motion. This pdf along with the pdf that describes the object motion form the system model of a Particle Filter, which is used to predict the most probable object locations between consecutive frames. Taking into account several camera motion candidates allows to handle challenging situations, such as the presence of multiple independent moving objects, and backgrounds

978-0-7695-3718-4/09 $25.00 © 2009 IEEE
DOI 10.1109/AVSS.2009.89

highly affected by the aperture problem. Each hypothesis about the object location is evaluated or weighted by the measurement model of the Particle Filter. This characterizes the target object by means of a spatiogram [3], which is a color histogram extended with structural information. The posterior pdf of the object location is computed by comparing the object spatiogram with those ones corresponding to each predicted location. The final estimation of the object location is finally obtained by means of the Maximum A Posteriori (MAP) estimator.

The rest of the paper is organized as follows: in Sec. 2 the Particle Filter framework for tracking with moving cameras is explained. The system model describing the camera and object dynamics is introduced in Sec. 3. The measurement model based on spatiograms is presented in Sec. 4. Experimental results are shown in Sec. 5, and conclusions are drawn in Sec. 6.

2. Particle Filter Framework for Tracking

The Particle Filter framework [2] for tracking aims to estimate the state of a target 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 contains all the relevant information about the object for the tracking purpose. In the present work, $\mathbf{x}_k$ encodes the kinematic and the geometric information, given by

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

where $\mathbf{l}_k = [l^x_k, l^y_k, 1]^\top$ is the vector of homogeneous spatial coordinates of the object, $\dot{\mathbf{l}}_k = [\dot{l}^x_k, \dot{l}^y_k]^\top$ is the velocity information, and $\mathbf{s}_k = [s^M_k, s^m_k, s^\theta_k]^\top$ respectively contains the mayor axis, the minor axis, and the orientation of the ellipse that encloses the target object.

Noisy measurements $\mathbf{z}_k$ are represented by a vector of observations that is related to the object information contained in the state vector. These observations are the HSV color channels of the image data: $\mathbf{z}_k = [\mathbf{H}_k, \mathbf{S}_k, \mathbf{V}_k]^\top$.

Instead of computing deterministically the state of the object, the Particle Filter algorithm recursively calculates some degree of belief in the state $\mathbf{x}_k$ at time k, using all the available information, including the set of measurements $\mathbf{z}_{1:k} = \{\mathbf{z}_i, i = 1, ..., k\}$ up to time k. Thus, the tracking task can be formulated as the estimation of the posterior pdf $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ of the state of the object. This pdf is approximated at each time step by a set of N_s weighted random samples

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) \approx W^{-1}_{N_s} \sum_{i=1}^{N_S} w^i_k \delta(\mathbf{x}_k - \mathbf{x}^i_k), \tag{2}$$

where the function $\delta(x)$ is the Kronecker's delta, and $W_{N_s} = \sum_{i=1}^{N_S} w^i_k$ is a normalization factor. As the number of samples becomes very large, this approximation becomes equivalent to the true posterior pdf.

Under certain assumptions [2], both samples, $\mathbf{x}^i_k$, and weights, w^i_k, can be recursively computed by means of the principle of the importance sampling, which formulates w^i_k as

$$w^i_k = w^i_{k-1} \frac{p(\mathbf{z}_k|\mathbf{x}^i_k)p(\mathbf{x}^i_k|\mathbf{x}^i_{k-1})}{q(\mathbf{x^i}_k|\mathbf{x}^i_{k-1}, \mathbf{z}_k)}. \tag{3}$$

The state transition probability $p(\mathbf{x}^i_k|\mathbf{x}^i_{k-1})$ is defined by the system model, which predicts the temporal evolution of the state of the object according to the dynamics of the camera and the target object.

The observation likelihood $p(\mathbf{z}_k|\mathbf{x}^i_k)$, defined by the measurement model, updates the predicted $p(\mathbf{x}^i_k|\mathbf{x}^i_{k-1})$ with current image data $\mathbf{z}_k = [\mathbf{H}_k, \mathbf{S}_k, \mathbf{V}_k]^\top$.

The pdf $q(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_k)$ is the importance sampling function, used to draw the samples $\mathbf{x}^i_k$. The pdf $q(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{z}_k)$ is usually approximated by $p(\mathbf{x}^i_k|\mathbf{x}^i_{k-1})$, what simplifies the Eq 3 to

$$w^i_k = w^i_{k-1} p(\mathbf{z}_k|\mathbf{x}^i_k). \tag{4}$$

Once computed the posterior pdf $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ as stated above, the estimation of the state vector of the object $\widehat{\mathbf{x}}_k$ at time step k is obtained by means of the MAP estimator, given by

$$\widehat{\mathbf{x}}_k = MAP(\mathbf{x}_k) = \arg\max_{\mathbf{x}_k} p(\mathbf{x}_k|\mathbf{z}_{1:k}). \tag{5}$$

The following sections respectively describe the system and the measurement models, used to update the weights w^i_k in each time step according to Eq 3.

3. System model

The temporal evolution of the state of the object depends on both target object dynamics and camera dynamics. A moving camera induces a global motion in the image that affects the expected location of the moving object. This fact is represented in the Fig. 1. This dual dependency is encoded in the system model of the state of the object as

$$\mathbf{x}_k = \mathbf{B}_{k-1}(\mathbf{A}\mathbf{x}_{k-1} + \mathbf{v}_{k-1}), \tag{6}$$

where $(\mathbf{A}\mathbf{x}_{k-1} + \mathbf{v}_{k-1})$ describes the dynamics of the object, and $\mathbf{B}_{k-1}$ represents the camera dynamics. The matrix $\mathbf{A}$ represents a linear model of constant velocity and size given by

$$\mathbf{A} = \left[\begin{array}{c|c|c|c} \mathbf{I}_2 & \mathbf{0}_{2\times1} & \mathbf{I}_2 & \mathbf{0}_{2\times3} \\ \hline \mathbf{0}_{6\times2} & \multicolumn{3}{c}{\mathbf{I}_6} \end{array} \right] \tag{7}$$

where $\mathbf{I}_2$ and $\mathbf{I}_6$ are respectively identity matrices of size 2×2 and 6×6, and $\mathbf{0}_{2\times1}$, $\mathbf{0}_{2\times3}$, and $\mathbf{0}_{6\times2}$ are respectively zeros matrices of size 2×1, 2×3, and 6×2.

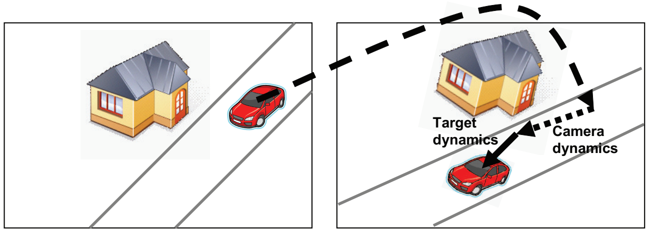

Figure 1. The object location between consecutive time steps depends on both camera dynamics and target object dynamics.

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

The matrix $\mathbf{B}_{k-1}$ representing the camera dynamics is defined by

$$\mathbf{B}_{k-1} = \left[\begin{array}{c|c} \mathbf{g}_{k-1} & \mathbf{0}_{3\times5} \\ \hline \mathbf{0}_{5\times3} & \mathbf{I}_5 \end{array} \right] \tag{8}$$

where $\mathbf{I}_5$ is an identity matrix of size 5×5, $\mathbf{0}_{3\times5}$ is a zero matrix of size 3×5, and $\mathbf{0}_{5\times3}$ is a zero matrix of size 5×3. The variable $\mathbf{g}_{k-1}$ is an independent stochastic process that represents the camera motion as a 2D affine transformation. This geometric transformation is a satisfactory approximation of the projective camera model, provided that the depth relief of the objects in the scene is small enough compared to the average depth, and the field of view is also small. At time step k, the output of $\mathbf{g}_{k-1}$ is a random variable whose pdf expresses the probability that the camera motion can be described by a certain affine transformation. Despite the pdf of $\mathbf{g}_{k-1}$ is unknown, it is possible to draw samples (i.e. affine transformation candidates) from $\mathbf{g}_{k-1}$ using the importance sampling principle. The drawing process starts computing correspondences between features detected in consecutive frames. For this purpose the SIFT algorithm [5] has been used, which is able to obtain reliable correspondences thanks to its robustness to noise and 3D view point changes, and its invariance to changes in scale, rotation and illumination. Then, the affine transformation candidates are drawn by randomly selecting combinations of three correspondences, since it is the minimum number of point pairs to infer an affine transformation. The set of combinations of three correspondences induces a subspace of the most probable affine transformations. However, this subspace can contain erroneous transformations due to the presence of outliers, caused by independent moving objects, the appearance/disappearance of image regions, and the aperture problem. RANSAC is a robust statistical algorithm [7] that can be used to compute the minimum number of affine transformations N_{AT} that ensures with a probability p_s that at least one is true, since it has been computed from a combination of correspondences without outliers. The expression of N_{AT} is

$$N_{AT} = \frac{\log(1 - p_s)}{\log(1 - (1 - \varepsilon)^3)} \tag{9}$$

where ε is the expected maximum fraction of outliers.

According to this, the minimum number of particles N_S used in the Particle Filter should be higher than N_{AT} to ensures that the camera motion is correctly represented at least by one sample.

4. Measurement model

The measurement model uses a spatiogram [3] to model the appearance of the object. Spatiograms are histograms augmented with spatial means and covariances to capture a richer description of the object. The spatiogram of a region belonging to the intensity image $\mathbf{I}$ is defined as a vector whose components are given by $h(b) = [n_b, \mu_b, \Sigma_b], b = 1, ..., B$, where n_b is the number of pixels whose values belong to that of the b^{th} bin, and μ_b and Σ_b are respectively the mean vector and covariance matrix of the spatial coordinates of the pixels contributing to the b^{th} bin. The spatiogram of an HSV image is computed in a similar way extending the unidimensional space of intensity pixel values to the tridimensional space of HSV pixel values.

The similarity between two spatiograms h and $h'$is computed by a weighted version of the Bhattacharyya coefficient, given by

$$\rho_W(h, h') = \Sigma_{b=1}^{B} w_b \rho(n_b, n'_b), \tag{10}$$

where $\rho(n_b, n'_b) = \sqrt{n_b n'_b}$, and the weights are defined by $w_b = N(\mu_b, \mu'_b, \Sigma'_b) N(\mu'_b, \mu_b, \Sigma_b)$, being $N(n, \mu, \Sigma)$ a multivariate Gaussian function evaluated at n.

Then, the measurement model evaluates the probability that the spatiogram $h(\mathbf{x}_k^i)$ of a candidate region defined by $\mathbf{x}_k^i$ is similar to the spatiogram of the object $h(\widehat{\mathbf{x}}_{k-1})$ by means of the expression

$$p(\mathbf{z}_k|\mathbf{x}_k^i) = \frac{1}{\sqrt{2\pi}\sigma} \exp - \left(\frac{d_W(h(\mathbf{x}_k^i), h(\widehat{\mathbf{x}}_{k-1}))^2}{2\sigma^2} \right) \tag{11}$$

where $d_W = \sqrt{1 - \rho_W}$ is the Bhattacharyya distance, and σ is the expected temporal variation of the Bhattacharyya distance due to temporal disturbances of the object appearance.

The object model, i.e. the spatiogram, is updated in each instant by the image region defined by the MAP estimation of the state $\widehat{\mathbf{x}}_k$.

5. Results

The presented visual tracking strategy for moving cameras is tested in two challenging situations, consisting in a chase of a object at high speed, where the camera that acquires images is mounted in a moving platform. This kind of sequences are ideal for testing the proposed tracking algorithm because they involve strong ego motions, and multiple moving independent objects, which can deviate the camera ego motion estimation. In both situations the posterior pdf of the state of the object has been approximated by 50 samples or particles.

The first sequence has been acquired by a camera mounted in a police car that is chasing a motorbike. The main challenges are the strong ego motion and the reduced set of appropriate regions to compute reliable correspondences. The major part of the image is composed by low textured regions in which the quality of correspondences is very poor due to the aperture problem. The proposed model for the camera dynamics, that takes into account multiple hypotheses, allows to manage this situation, and thus to perform successfully the tracking of the target object, as shown in Fig. 2. The first row of frames corresponds to two non consecutive time steps of the sequence, where the tracked target object (the motorbike) is enclosed by a white ellipse. The second row shows the same frames along with a set of ellipses that represents the set object state samples considering only the camera dynamics. Note that ellipses are not exactly located on the target object, since only the camera dynamics has been taken into account. Using the object dynamics to modify the object location candidates, ellipses are finally located on the image region of the target object, as shown in the third row. The combination of the target object and camera dynamics allows to satisfactorily propagate the posterior pdf of the state of the object.

The second sequence has been acquired from a camera mounted on a helicopter. The chased object in this case is a red car. In addition to the strong ego-motion, there are several independent moving objects (other cars) that makes more challenging the camera motion estimation, since the correspondences between the moving objects can induce false camera motions. Fig. 3 shows three rows of frames with the same disposition and meaning as in Fig. 2. Note that ellipses of the second and third rows are distributed along the direction of the highway because the correspondences between SIFT features belonging to cars induce several false camera motion candidates in such direction. In spite of this drawback, the tracking is successfully accomplished as observed in the first row, where the white ellipse encloses the tracked object. Notice how this strategy allows to discard as possible candidate the other red car (second column of frames) marked with a white X, since there is no camera nor object motion that supports that region. This avoids that the tracking process can be distracted by similar objects.

The performance of the proposed tracking algorithm has been compared with two tracking techniques described in [6] and [8]. The main differences with the presented approach are that in [6] the camera ego-motion is not compensated, while in [8] it is compensated, but using only one affine transformation instead of several ones. All of three strategies use a Particle Filter framework to perform the tracking. In order to appropriately compare the three tracking algorithms, the same set of parameters has been used to tune the Particle Filters. The dataset used to perform the comparison is composed by 11 videos with challenging situations: strong ego-motion, changes in illumination, variations of the object appearance and size, and occlusions. The whole dataset, along with the tracking results, can be downloaded from the website: `http://www.gti.ssr.upm.es/paper/avss09/`. The comparison is based on the following criteria: number of times that the tracked object has been lost. Each time that the tracked object is lost, the corresponding tracking algorithm is initiated with a correct object detection from the same frame where the object was lost. Table 1 shows the comparison of the three tracking algorithms, where it can be appreciated that proposed algorithm is quite superior, due to the other approaches can not satisfactorily handle the ego-motion. On the other hand, it can be observed that, independently of the algorithm, the results obtained by the videos “motorbike1-6.avi” are worse than the rest. The reason is that the size of the tracked object is very small, less than one hundred pixels, and therefore the object can not be robustly represented by a spatiogram.

Video	#L alg. 1	#L alg. 2	#L alg. 3
redcar1.avi	0	0	0
redcar2.avi	1	4	9
person1.avi	0	3	7
motorbike1.avi	4	13	24
motorbike2.avi	1	3	6
motorbike3.avi	2	8	19
motorbike4.avi	5	16	26
motorbike5.avi	8	23	37
bluecar1.avi	0	1	3
bluecar2.avi	0	2	5
bluecar3.avi	1	3	6

Table 1. Comparison between the three tracking algorithms using the criteria of number of times that the tracked object has been lost (#L), where alg. 1, alg. 2, and alg. 3 refer the proposed method, the Particle Filter approach with compensation, and the Particle Filter approach without compensation, respectively.

6. Conclusions

The presented visual tracking strategy is able to track an object in complex situations involving strong camera ego motion and multiple independent moving objects. The

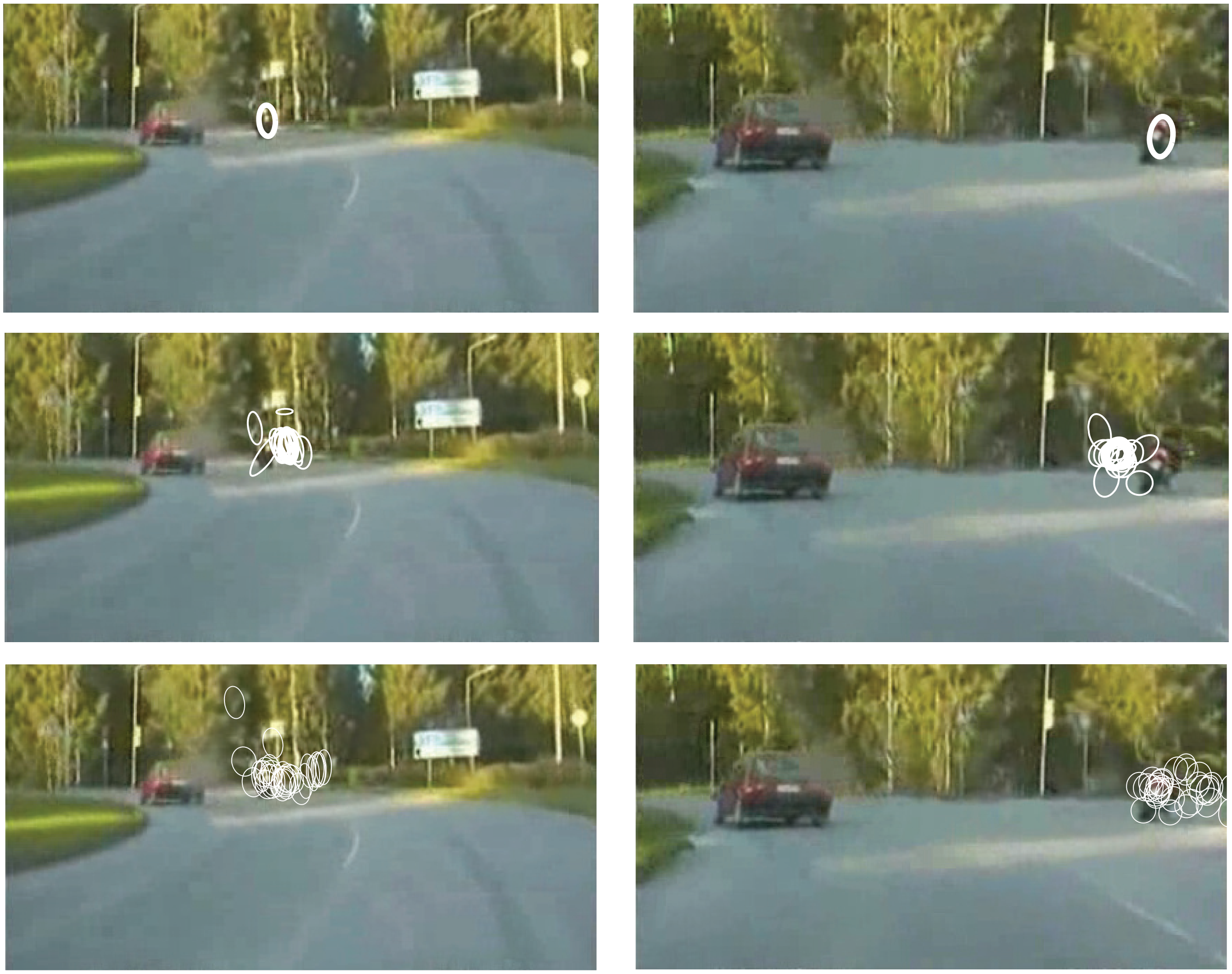

Figure 2. Tracking results in a motorbike chase from a camera mounted on a car. The first row shows the estimation of the object state represented by a white ellipse for two different frames. The second row shows candidates of the object state taking only into account the camera dynamics. The third row shows the modified candidates of the object state using also the object dynamics.

tracking problem is modeled by a Particle Filter which uses an advanced system model that takes into account not only the object dynamics, but also the camera dynamics. The main novelty arises from modeling the camera dynamics by an independent stochastic process. Thus, in each time step the camera dynamics is represented by a discrete pdf that encodes the most probable global affine transformations between consecutive frames. This dramatically improves the robustness of the tracking, especially in those situations where the camera motion can not be obtained accurately (sequences highly affected by the aperture problem) and/or there are several possible camera motions (sequences containing several independent moving objects). In this sense, the proposed approach outperforms the existing techniques such us [9, 1, 8], which perform the tracking considering only one hypothesis of camera motion. As a consequence of this, the tracking process may fail if the estimated camera motion does not correspond with the true one. Experimental results support the previous assertion, showing the high performance of the presented tracking approach in challenging situations.

7. Acknowledgements

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).

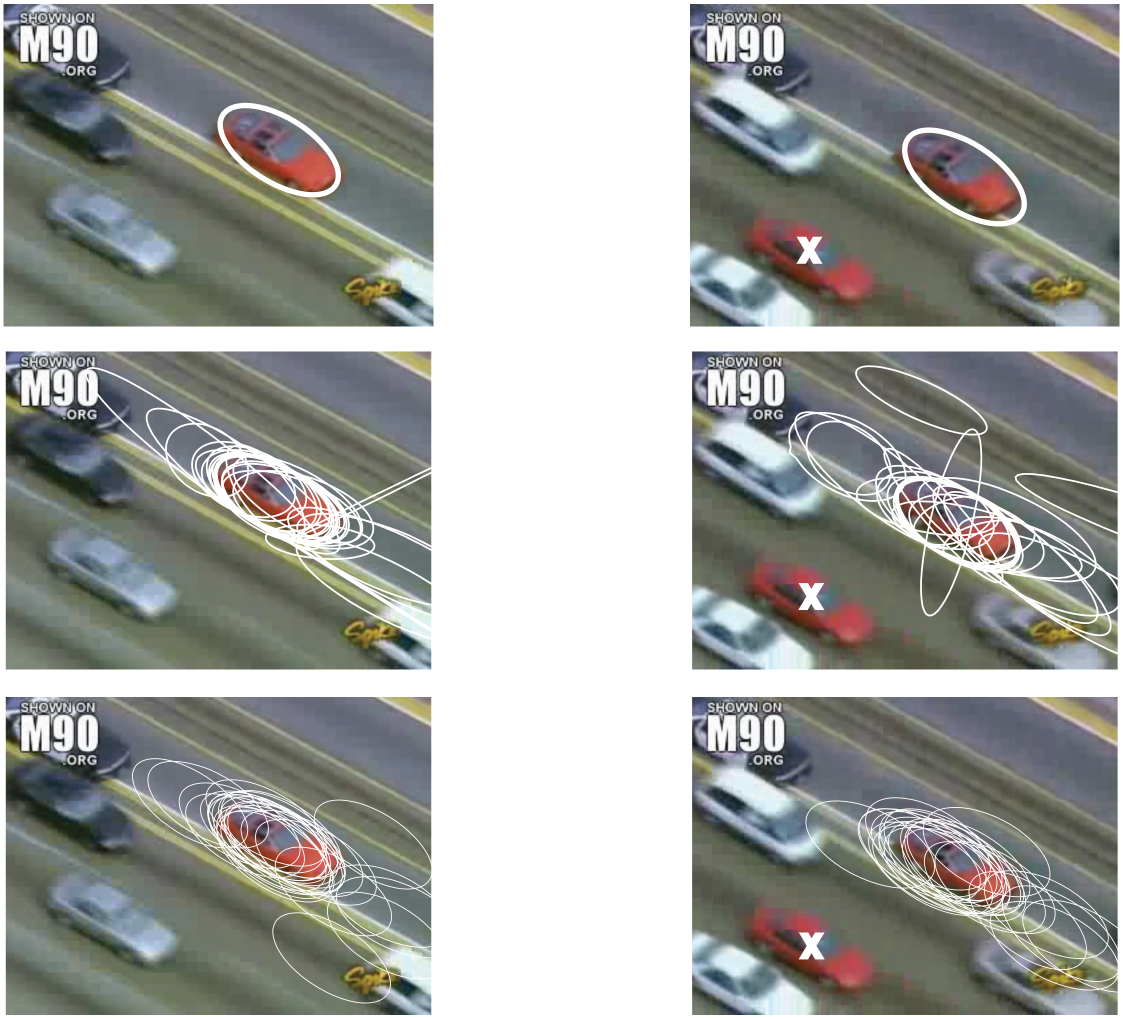

Figure 3. Tracking results in a car chase from a camera mounted on a helicopter. The first row shows the estimations of the object state represented by a white ellipse for two different frames. The second row shows candidates of the object state taking only into account the camera dynamics. The third row shows the modified candidates of the object state using also the object dynamics. The white X in the second column indicates a similar object (other red car) that could be easily mistaken for the target object.

References

[1] S. Ali and M. Shah. Cocoa: tracking in aerial imagery. In *SPIE Proc. Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems and Applications III*, volume 6209, 2006. 1, 5

[2] S. Arulampalam, S. Maskell, and N. Gordon. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50:174–188, 2002. 2

[3] S. Birchfield and S. Rangarajan. Spatial histograms for region-based tracking. *ETRI Journal*, 29(5):697–699, 2007. 2, 3

[4] J. Domke and Y. Aloimonos. A probabilistic notion of correspondence and the epipolar constraint. In *Proc. of the Third International Symposium on 3D Data Processing, Visualization, and Transmission*, pages 41–48, 2006. 1

[5] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Computer Vision*, 60(2):91–110, 2004. 3

[6] K. Nummiaro, E. Koller-Meierb, and L. V. Gool. An adaptive color-based particle filter. *Image and Vision Computing*, 21(1):99–110, 2003. 4

[7] P. Torr and D. Murray. The development and comparison of robust methodsfor estimating the fundamental matrix. *Int. J. Computer Vision*, 24(3):271–300, 1997. 3

[8] V. Venkataraman, G. Fan, and X. Fan. Target tracking with online feature selection in flir imagery. *IEEE Proc. CVPR*, pages 1–8, 2007. 1, 4, 5

[9] A. Yilmaz, K. Shafique, and M. Shah. Target tracking in airborne forward looking infrared imagery. *Image and Vision Computing*, 21(7):623–635, 2003. 1, 5

[10] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, 2003. 1