Camera Localization Using Trajectories and Maps

Raúl Mohedano, Andrea Cavallaro, and Narciso García

Abstract—We propose a new Bayesian framework for automatically determining the position (location and orientation) of an uncalibrated camera using the observations of moving objects and a schematic map of the passable areas of the environment. Our approach takes advantage of static and dynamic information on the scene structures through prior probability distributions for object dynamics. The proposed approach restricts plausible positions where the sensor can be located while taking into account the inherent ambiguity of the given setting. The proposed framework samples from the posterior probability distribution for the camera position via data driven MCMC, guided by an initial geometric analysis that restricts the search space. A Kullback-Leibler divergence analysis is then used that yields the final camera position estimate, while explicitly isolating ambiguous settings. The proposed approach is evaluated in synthetic and real environments, showing its satisfactory performance in both ambiguous and unambiguous settings.

Index Terms—Vision and scene understanding, camera calibration, Markov processes, tracking

1 Introduction

Visual-based camera self-positioning is fundamental for automatic vehicle guidance, aerial imaging, photogrammetry and calibration of previously placed cameras (e.g., CCTV networks). Existing methods use general geometric/dynamic assumptions without any knowledge of the site being observed, or use specific environmental knowledge (e.g., in the form of a map). Visual observation of invariants that are independent from the specific camera location can be used for camera position estimation [1], [2]: for instance, the observation of stars [3], vertical lines and vanishing points of architectural structures [4] can be used to infer sensor orientation, without providing camera location estimations. Alternatively, trajectory models can also be used for extrinsic parameter calibration of camera networks [5], [6], [7], [8] and for temporal synchronization [9], [10]. Regular object dynamics such as polynomials of known degree [11] and probabilistic linear Markov models [12], [13] are generally assumed to allow the estimation of *relative* sensor positions from reconstructed trajectories inside and outside the field of view (FoV) of the cameras.

Most works that use environmental knowledge assume prior information obtained offline. Only certain Simultaneous Localization And Mapping (SLAM) systems build online 3D descriptions for sequential re-localization [14], [15]. There are however SLAM approaches that use 3D scene models to match the inferred 3D models to improve the robustness and accuracy of the resulting positioning [16], [17]. Direct image registration between the camera's FoV and an aerial/satellite map of the site under study [18] is a valid option with planar environments or aerial cameras. Matching approximate observations of environmental landmarks of the ground-plane of the scene and their actual positions on the street map assumes manual user interaction [19], or automatic feature extraction from both satellite and camera views [20]. A more versatile approach but less accurate and more demanding in terms of map size and processing uses geo-referenced street view data sets (such as Google Street View or OpenStreetMap) for absolute camera positioning [21], [22].

In this paper, we present a Bayesian framework for inferring, from the observed trajectories of moving objects and a schematic map of the scene indicating passable areas of the environment, the absolute location and orientation (i.e., position) of a fixed camera whose view of the dominant ground-plane of the scene is assumed metrically rectified. We consider prior probability distributions for object trajectories taking into account the dynamics encouraged by the structures defined in the map of the specific scene, and compare them to the 2D tracks to obtain the posterior probability distribution for the camera position. This posterior is analyzed using Monte Carlo probabilistic sampling and Kullback-Liebler divergence-based clustering to obtain the final estimation for the given setting. It is important to note that, unlike existing visual positioning systems that do not perform any reliability analysis [11], [12], [13], our proposal provides not only an estimation of the absolute camera position but also the inherent uncertainty associated to the setting when it is not unambiguous. Moreover, to the best of our knowledge, no existing approaches have explored the use of the moving objects observed by the camera for *absolute* positioning. Indeed, the use of trajectories have only been reported for sequential absolute positioning of moving sensors, and consist in the comparison between the reconstructed motion of the sensor, performed either using visual

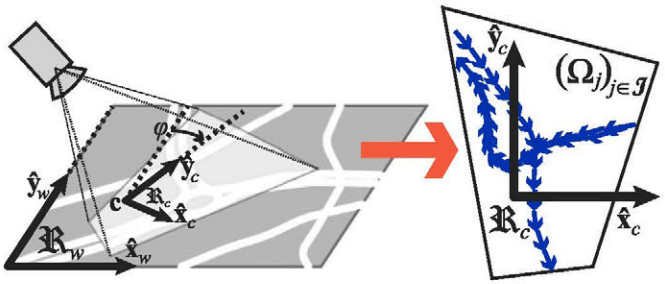

Fig. 1. Geometric meaning of the parameters $\mathrm{E}=(\mathbf{c},\varphi)$, and description of the relation between camera FoV (via known local camera-to-ground homography) and local and absolute (map-related) frames of reference.

SLAM [23] or even GPS sensing [24], [25], and a simple schematic plan containing only the passable areas of the scene. These approaches are thus limited to moving sensors, and operate only using one trajectory.

The remainder of the paper is organized as follows. Section 2 defines the estimation task. Section 3 presents the proposed approach, which is divided into preprocessing and principal module, detailed in Sections 4 and 5, respectively. The effectiveness of the proposed approach is analyzed with traffic data in Section 6, and the conclusions on the work and an outlook on its potential extensions are discussed in Section 7.

2 Problem Statement

Let us assume a metric rectification homography providing a virtual bird's-eye-view of the observation of a camera of the ground-plane containing the passable regions of the scene [26], [27]. Let the positions $\mathbf{p}_c \in \mathbb{R}^2$ observed on the camera image plane be referred to a 2D local frame of reference $\mathfrak{R}_c$ of the ground plane, corresponding to the said metric rectification of the camera's FoV (Fig. 1). $\mathfrak{R}_c$ can be related point-wise to the absolute ground-plane frame $\mathfrak{R}_w$ by means of the orientation-preserving isometry

$$m(\mathbf{p}_c;\mathbf{c},\varphi) = \mathrm{R}(\varphi)\,\mathbf{p}_c + \mathbf{c}\,, \tag{1}$$

where $\mathrm{R}(\varphi)$ is the rotation matrix of angle φ. The isometry (1) is described by two camera parameters: $\mathbf{c}=(x_c,y_c)\in\mathbb{R}^2$, the translation of the camera local coordinate system $\mathfrak{R}_c$ with respect to the global frame $\mathfrak{R}_w$, and $\varphi\in[0,2\pi)$, the angle between both systems (it is actually a parameterization of $\mathrm{SO}(2)$, whose "cyclic" nature must be taken into account in subsequent considerations). For this reason, we simply encode the extrinsic parameters of the camera using the pair $\mathrm{E}=(\mathbf{c},\varphi)$.

Let Ω be a set of N_Ω 2D tracks, $\Omega=(\Omega_j)_{j\in\mathcal{J}}$, $\mathcal{J}=\{1,\ldots,N_\Omega\}$. Each track is a sequence of time-stamped 2D points $\Omega_j=(\omega_j^t)_{t\in T(\Omega_j)}$ (with $\omega_j^t\in\mathbb{R}^2$) during the interval $T(\Omega_j)$, referred to the camera local frame of reference $\mathfrak{R}_c$.

Our aim is to infer the most plausible hypothesis for the absolute positional parameters E (2D location and 1D orientation) from the set Ω of N_Ω 2D tracks and an available map M encoding the influence of the environment on object dynamics. The map affects the characteristics of object routes across the environment, and routes determine the observations Ω_j: this transitive relation is the key to solve this estimation problem.

3 Overview of the Proposed Approach

3.1 Scene Map M Definition

Let the information about the scene be encoded in the input structure M (Fig. 2) reflecting the restrictions imposed by the scene layout on object trajectories. We use a binary mask B_M for the delimitation of passable areas.

The capability to generate plausible path hypotheses requires a mechanism for selecting high-level paths linking two areas of the map, namely entry and exit zones. M is therefore defined using a dual representation consisting in the separation of passable regions into different units or nodes. Nodes are of three types: segments ($\mathcal{S}$), crossroads ($\mathcal{C}$), and gates ($\mathcal{G}$). $\mathcal{S}$ nodes are portions of the map linking two $\mathcal{C}$ or $\mathcal{G}$ nodes, $\mathcal{C}$ nodes represent regions where two or more $\mathcal{S}$ nodes converge, and $\mathcal{G}$ nodes indicate those regions where objects can enter/exit the scene. $\mathcal{G}$ nodes are distinguished between entry (appearance) and exit (disappearance), denoted respectively by $\mathcal{G}_A$ and $\mathcal{G}_D$. We denote by $\mathcal{M}$ the whole set of nodes $\mathfrak{n}$ of the map. Using this division of the scene, high-level paths between pairs of regions can be expressed as ordered sequences of nodes.

Additionally, we consider a connectivity matrix C_M indicating the direct vicinity between pairs of nodes in the structure M. This binary matrix expresses implicitly the direction(s) each node of the system can be traversed in. C_M is, by construction, sparse: this characteristic allows the use of efficient graph-search algorithms for

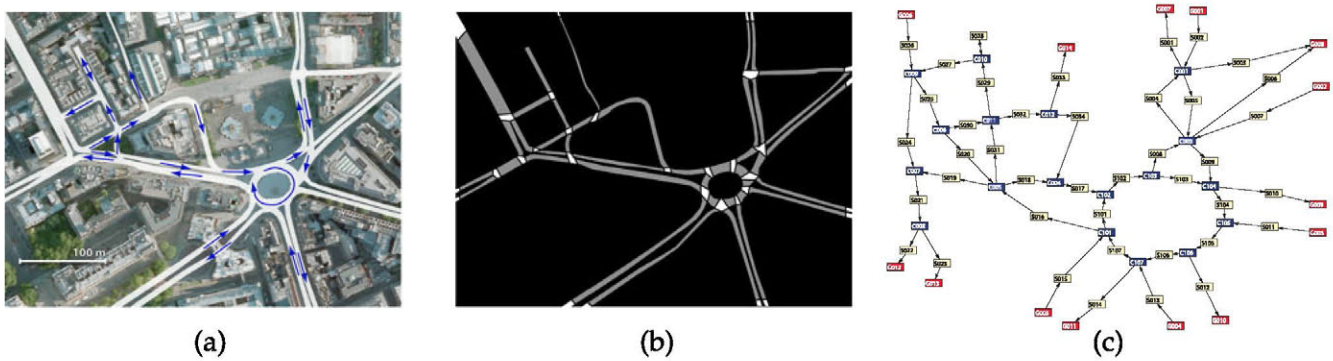

Fig. 2. Description of the map structure M. (a) Environment (Copyright 2012 Google-Map data). (b) Individual binary masks $\mathrm{B}_\mathfrak{n}$ for all the nodes composing M ($\mathcal{S}$ nodes in grey, $\mathcal{C}$ and $\mathcal{G}$ in white). (c) Connectivity graph ($\mathcal{S}$ in yellow, $\mathcal{C}$ in blue and $\mathcal{G}$ in red).

discovering high-level paths linking two areas of the map such as the classical k-shortest path search algorithm by Yen [28] or other equivalents [29].

To include spatial characteristics, each node $\mathfrak{n}$ has an associated binary mask $\mathrm{B}_{\mathfrak{n}}$ indicating its extent, and such that, ideally, $\mathrm{B}_{\mathfrak{n}} \cap \mathrm{B}_{\mathfrak{n}'} = \emptyset$, $\forall \mathfrak{n}, \mathfrak{n}' \in \mathcal{M}$ with $\mathfrak{n} \neq \mathfrak{n}'$, and $\mathrm{B}_{\mathrm{M}} = \bigcup_{\forall \mathfrak{n} \in \mathcal{M}} \mathrm{B}_{\mathfrak{n}}$. Additionally, each $\mathcal{G}_A$ node is described as a point $\mathbf{b}_{\mathfrak{n}} \in \mathbb{R}^2$ indicating its corresponding expected object entry position.

3.2 Proposed Bayesian Formulation

The inference of E is based on two main sources of information: the partial locally observed trajectories $(\Omega_j)_{j \in \mathcal{J}}$ and the map M. Their influence has a considerable associated uncertainty, which can be modeled statistically. We analyze the camera positional parameters E and their associated uncertainty by expressing the posterior distribution $p(\mathrm{E} \,|\, \Omega, \mathrm{M})$ from both sources. However, E and the observation Ω_j of a certain object are not directly related: the observed track Ω_j will be determined by E and *by the absolute trajectory* R_j *of the actual underlying object which caused the observation.* For this reason, we introduce explicitly object trajectories $\mathrm{R} = (\mathrm{R}_j)_{j \in \mathcal{J}}$ (with R_j as defined in Section 5.1) as auxiliary variables to indirectly obtain the posterior distribution for E: defining first the posterior joint distribution $p\big(\mathrm{E}, \mathrm{R} \,|\, \Omega, \mathrm{M}\big)$, and then marginalizing it over the space of possible object routes to obtain

$$p(\mathrm{E} \,|\, \Omega, \mathrm{M}) \;=\; \int p(\mathrm{E}, \mathrm{R} \,|\, \Omega, \mathrm{M}) \, \mathrm{dR}\,, \tag{2}$$

where the integral should be considered mere notation. The use of the indirect expression (2) is motivated by the fact that the joint distribution allows a satisfactory decomposition in terms of the two previously discussed sources of information as, via Bayes' rule, we can write

$$p(\mathrm{E}, \mathrm{R} \,|\, \Omega, \mathrm{M}) \;\propto\; p\big(\Omega \,|\, \mathrm{E}, \mathrm{R}, \mathrm{M}\big)\, p\big(\mathrm{E}, \mathrm{R} \,|\, \mathrm{M}\big). \tag{3}$$

The first factor on the right-hand side, the *observation* or *likelihood model*, represents the probability distribution of the 2D tracks Ω for given E and R. The second factor, the *prior distribution*, encodes the information on the unknowns before any experimental data have been observed. Their definitions are detailed in Section 5.1.

This indirect definition of the posterior distribution $p\big(\mathrm{E} \,|\, \Omega, \mathrm{M}\big)$ does not allow an analytic study of its main characteristics. For this reason, we construct an empirical distribution from a set of U samples $\{\overline{\overline{\mathrm{E}}}^{(u)}\}_{u=1}^{U}$ drawn from it, obtained using Markov chain Monte Carlo (MCMC) methods [30].

Since the observed tracks could be consistent with multiple camera positions in the map (e.g., straight passable regions that do not present enough distinctive features), the estimation of the camera position may present an inherent ambiguity. We detect these ambiguous settings by analyzing the sample-based representation $\{\overline{\overline{\mathrm{E}}}^{(u)}\}_{u=1}^{U}$ of $p\big(\mathrm{E} \,|\, \Omega, \mathrm{M}\big)$. This analysis, performed using an adaptation of the K-adventurers algorithm [31], allows us to infer the set $\{\overline{\mathrm{E}}^{(k)}\}_{k=1}^{K}$ of $K < U$ distinct approximate camera position hypotheses that best approximates the sampled distribution in terms of the Kullback-Leibler (KL) divergence [32]. This generates an estimate of the relative plausibility of the retained hypotheses that quantifies the ambiguity of the given setting.

The generation of the samples $\{\overline{\overline{\mathrm{E}}}^{(u)}\}_{u=1}^{U}$ via MCMC requires at each iteration the evaluation of the posterior probability for the proposed camera hypothesis. However, the marginalization process described in (2) cannot be performed analytically. For this reason we design a sequential Monte Carlo algorithm based on importance sampling (Appendix A). Although efficient, this process impacts the computational cost of camera hypothesis evaluation. To overcome this problem, we take two additional steps. First, we simplify the set of available tracks using unsupervised trajectory clustering [33] that generates a reduced set of representative tracks. Second, we define a *preprocessing module* aimed at guiding the search within the solution space and thus compensate for the fact that $p(\mathrm{E} \,|\, \Omega, \mathrm{M})$ presents, in general, a considerable number of sparsely distributed modes. This module estimates a proposal distribution $q(\mathrm{E} \,|\, \Omega, \mathrm{M})$ by checking the purely geometric consistency between observed object positions and passable regions of the map. The *principal module* concerns the MCMC sampling process itself, guided by this proposal $q\big(\mathrm{E} \,|\, \Omega, \mathrm{M}\big)$, and the Kullback-Leibler divergence analysis (Fig. 3).

4 Preprocessing: Geometric Analysis

The aim of the geometric analysis module is the generation of a proposal distribution density $q(\mathrm{E} \,|\, \Omega, \mathrm{M})$ (from now on, $q(\mathrm{E} \,|\, \mathrm{M})$) expressing the plausibility of each camera position according to the purely geometric consistency between the observed 2D tracks and the map.

Let us assume a positive function $f(\mathrm{E}; \Omega, \mathrm{M})$, evaluable at each E and measuring the fit of point-wise observations and binary mask B_{M}, such that $q(\mathrm{E} \,|\, \mathrm{M}) \propto f(\mathrm{E}; \Omega, \mathrm{M})$. We can use a different importance sampling distribution $z(\mathrm{E})$ for extracting samples from to obtain the set $\{\tilde{\mathrm{E}}^{(s)}, \tilde{w}^{(s)}\}_{s=1}^{S}$ of weighted samples, with $\tilde{\mathrm{E}}^{(s)} \sim z(\mathrm{E})$ and

$$\tilde{w}(\tilde{\mathrm{E}}^{(s)}) = \frac{q(\tilde{\mathrm{E}}^{(s)} | \mathrm{M})}{z(\tilde{\mathrm{E}}^{(s)})}.$$

These samples are used for approximating the desired density $q(\mathrm{E} \,|\, \mathrm{M})$ by the kernel-based function

$$\tilde{q}(\mathrm{E} | \mathrm{M}) = \frac{1}{S} \sum_{s=1}^{S} w(\tilde{\mathrm{E}}^{(s)})\, k(\mathrm{E} - \tilde{\mathrm{E}}^{(s)}),$$

where $k\big(\mathrm{E}\big)$ is a multidimensional kernel function (integrating one over its whole domain), defined over the continuous space $\mathbb{R}^2 \times \mathrm{SO}(2)$ (therefore, kernel density estimation techniques adapted to this partially cyclic space are used). This approximation is both operational and accurate, and converge as $S \to \infty$ to the convolution $q\big(\mathrm{E} \,|\, \mathrm{M}\big) * k\big(\mathrm{E}\big)$ (Appendix B). For simplicity, we use independent kernels for each dimension of E. The kernels of the two spatial dimensions have been chosen Gaussian. The kernel of the angular component has been considered Gaussian with truncated tails,

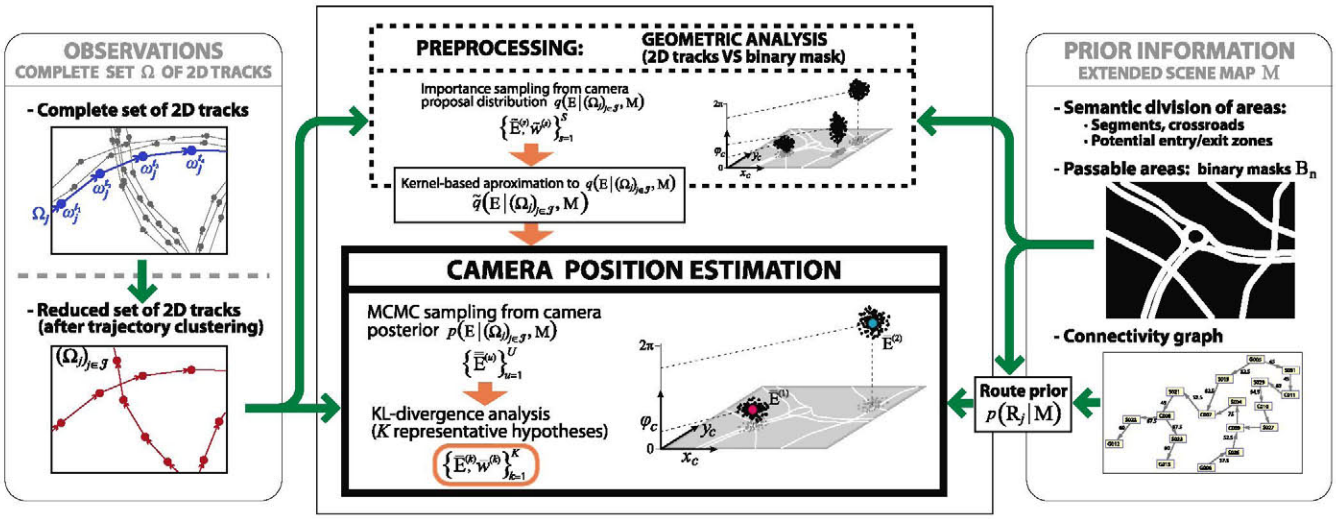

Fig. 3. System overview: preprocessing and principal modules, and input and output variables at each stage.

since angular kernel bandwidth must be clearly lower than 2π to avoid affecting excessively the shape of the obtained kernel-based approximation.

We define the importance distribution as

$$z(\mathrm{E}) = z(\mathbf{c}, \varphi) \equiv p(\mathbf{c} \,|\, \mathrm{M}) q(\varphi \,|\, \mathbf{c}, \mathrm{M}) \,, \tag{4}$$

where $p(\mathbf{c} \,|\, \mathrm{M})$ represents a certain prior distribution for the camera location aimed at increasing the rate of camera hypotheses generated in the proximity of the passable regions of the map (but with no direct relationship to the calculation of the posterior distribution (7)); $q(\varphi \,|\, \mathbf{c}, \mathrm{M})$ is expressly chosen so as to coincide with the conditional distribution of φ. The definition of $p(\mathbf{c} \,|\, \mathrm{M})$ is based on the assumption that cameras are placed in order to monitor moving object behaviors. Thus, assuming that the origin of the camera local coordinate system $\Re_c$ is close to the centroid of the transformed FoV onto the ground plane, locations $\mathbf{c}$ that are closer to the passable areas of the map should have a higher probability. This is fulfilled by the distribution

$$p(\mathbf{c} \,|\, \mathrm{M}) = \frac{1}{|\breve{\mathrm{B}}_\mathrm{M}|} \sum_{\breve{\mathbf{c}} \in \breve{\mathrm{B}}_\mathrm{M}} \frac{1}{2\pi\sigma_c^2} e^{\frac{-1}{2\sigma_c^2}(\mathbf{c}-\breve{\mathbf{c}})^T(\mathbf{c}-\breve{\mathbf{c}})} \,,$$

where $\breve{\mathrm{B}}_\mathrm{M}$ represents the set of map points corresponding to the center positions of the pixels indicated as passable in the binary mask of the scene and σ_c controls the extent around B_M.

The definition of the importance density (4) simplifies weight calculation into

$$\tilde{w}(\tilde{\mathrm{E}}^{(s)}) = \frac{q(\tilde{\mathbf{c}}^{(s)} \,|\, \mathrm{M})}{p(\tilde{\mathbf{c}}^{(s)} \,|\, \mathrm{M})} \propto \frac{\int_0^{2\pi} f(\tilde{\mathbf{c}}^{(s)}, \varphi; \Omega, \mathrm{M}) \, \mathrm{d}\varphi}{p(\tilde{\mathbf{c}}^{(s)} \,|\, \mathrm{M})} \,, \tag{5}$$

and allows hierarchical sampling for $\mathrm{E}^{(s)}$ by first extracting $\tilde{\mathbf{c}}^{(s)}$ from $p(\mathbf{c} \,|\, \mathrm{M})$ and using this $\tilde{\mathbf{c}}^{(s)}$ to apply integral transform sampling [34] on $q(\varphi \,|\, \mathbf{c}, \mathrm{M})$. The integral

$$Q(\varphi; \tilde{\mathbf{c}}^{(s)}) = \int_0^{\varphi} q(\xi \,|\, \tilde{\mathbf{c}}^{(s)}, \Omega, \mathrm{M}) \, \mathrm{d}\xi = \frac{\int_0^{\varphi} f(\tilde{\mathbf{c}}^{(s)}, \xi; \Omega, \mathrm{M}) \, \mathrm{d}\xi}{\int_0^{2\pi} f(\tilde{\mathbf{c}}^{(s)}, \zeta; \Omega, \mathrm{M}) \, \mathrm{d}\zeta} \,,$$

required for transform sampling is obtained using linear interpolation on $f(\tilde{\mathbf{c}}^{(s)}, \varphi; \Omega, \mathrm{M})$ (as a function of φ only) by evaluating this function for a relatively low number of orientations. This approximated integral is also used to set the numerator of (5) for weight calculation.

To illustrate the proposed sampling approach, valid for general positive functions $f(\mathrm{E}; \Omega, \mathrm{M})$, we propose the following $f(\mathrm{E}; \Omega, \mathrm{M}) = r^\gamma$, where $\gamma \gg 1$ and

$$r(\Omega; \mathrm{E}, \mathrm{B}_\mathrm{M}) = \frac{1}{\sum_{j \in \mathcal{J}} |\Omega_j|} \sum^{j \in \mathcal{J}} \sum_{t \in T(\Omega_j)} e^{\frac{1}{2\sigma_\Omega^2} d_\mathrm{M}^2(m(\boldsymbol{\omega}_j^t; \mathrm{E}))} \,,$$

where σ_Ω is the standard deviation of the observation noise (Section 5.1.1) and $d_\mathrm{M}(\cdot)$ is the Euclidean distance to the map mask B_M. The proposed consistency-checking function highlights hypotheses E that are in line with the map and the observations while avoiding penalizing mismatches due to observation noise.

5 Camera Position Estimation

The camera position estimation module analyzes $p(\mathrm{E} \,|\, \Omega, \mathrm{M})$, defined as the marginal of the factorized joint distribution. Its main characteristics will be captured by a set of K weighted camera positions $\{\overline{\mathrm{E}}^{(k)}, \overline{w}^{(k)}\}_{k=1}^K$ representing hypotheses and the inherent ambiguity of the setting. For this purpose, U samples $\{\overline{\overline{\mathrm{E}}}^{(u)}\}_{u=1}^U$ are drawn from the posterior $p(\mathrm{E} \,|\, \Omega, \mathrm{M})$ by MCMC sampling using the result of the preprocessing module as a proposal density. The set $\{\overline{\mathrm{E}}^{(k)}, \overline{w}^{(k)}\}_{k=1}^K$ is estimated by searching the subset of K samples in the generated sample set $\{\overline{\overline{\mathrm{E}}}^{(u)}\}_{u=1}^U$ that approximates best the empirical posterior distribution in terms of the Kullback-Leibler divergence.

5.1 Probability Model Definition

Because camera E and route parameters R are conditionally independent given the map, the joint posterior distribution (3) can be further factorized:

$$p(\mathrm{E},\mathrm{R}\,|\,\mathrm{M}) \;=\; p(\mathrm{E}\,|\,\mathrm{M})\,p(\mathrm{R}\,|\,\mathrm{M}). \tag{6}$$

This factorization allows us to naturally integrate different sources of information on the absolute camera location, such as GPS, Wi-Fi based analysis [35] or other sensor localization algorithms [12]. We assume a non-informative prior for E, discarding from now on the term $p(\mathrm{E}\,|\,\mathrm{M})$ in the expression of the posterior distribution.

We assume conditional independence between routes of different objects $(\mathrm{R}_j)_{j\in\mathcal{J}}$ given M. Additionally, observations Ω_j corresponding to different objects can be considered conditionally independent given the true routes R_j that generated them and independent from the map. This consideration, along with those previously given to (6), allows us to write the joint posterior distribution in (3) in terms of individual observation models and route priors as

$$p(\mathrm{E},\mathrm{R}\,|\,\Omega,\mathrm{M}) \propto \prod_{j\in\mathcal{J}} [\, p\big(\Omega_j\,|\,\mathrm{E},\mathrm{R}_j\big)\, p\big(\mathrm{R}_j\,|\,\mathrm{M}\big)\,]. \tag{7}$$

The specific definitions proposed for the two individual probability distributions retained in (7) involve the auxiliary variables R_j representing real object trajectories during their presence in the scene. Their definition has been chosen to reflect the environmental influence on objects: they "decide" which high-level path to follow across the scene (i.e., entry and exit regions) and the sequence of regions followed to link them. Moreover, they move locally so as not to violate the positional and dynamical restrictions imposed by its previous high-level choice. These considerations leads to a two-level representation

$$\mathrm{R}_j \;=\; \big(\,\mathcal{H}_j\,,\,T(\mathrm{R}_j)\,,\,\big(\mathbf{r}_j^t,\mathbf{v}_j^t\big)_{t\in T(\mathrm{R}_j)}\big),$$

which explicitly contains the high-level path $\mathcal{H}_j$ followed by the object across the map as well as the low-level description $(\,T(\mathrm{R}_j),(\mathbf{r}_j^t,\mathbf{v}_j^t)_{t\in T(\mathrm{R}_j)}\,)$ of the route itself. The usefulness of this definition is clarified in Section 5.1.2.

$\mathcal{H}_j$ is defined as a sequence of high-level nodes of the map M expressing the areas to be traversed to link an entry and exit node. As for the low-level description of the route, the movement of the objects, of continuous nature, is modeled as a discrete process with the same framerate[1] of the observed 2D tracks Ω_j. $T(\mathrm{R}_j)$ indicates the interval of consecutive time steps during which the considered object is present in the scene (which is, in principle, unknown), and the time-stamped pairs $(\mathbf{r}_j^t,\mathbf{v}_j^t)$ represent the position and velocity of the object at time t expressed with respect to the absolute frame $\Re_w$. Although neither $\mathcal{H}_j$ nor $\mathbf{v}_j^t$ are directly observable, they are included in this definition to ease object dynamics modeling.

5.1.1 Observation Model

Our definition for the observation model $p\big(\Omega_j\,|\,\mathrm{E},\mathrm{R}_j\big)$ of each individual object is inspired by [12] and considers only positive observations (i.e., the lack of observation is not modeled). Although we write $p\big(\Omega_j\,|\,\mathrm{E},\mathrm{R}_j\big)$ for clarity, the observation process is the noisy registration of the true positions of the jth object over time. Thus, it would be more appropriate to explicitly indicate that neither the high-level path $\mathcal{H}_j$ followed by the object nor its velocity are actually involved. So,

$$p\big(\Omega_j\,|\,\mathrm{E},\mathrm{R}_j\big) = p\big(\big(\boldsymbol{\omega}_j^t\big)_{t\in T(\Omega_j)}\,|\,\mathrm{E},T(\mathrm{R}_j),\big(\mathbf{r}_j^t\big)_{t\in T(\mathrm{R}_j)}\big).$$

We assume that all possible observation time spans $T(\Omega_j)$ such that $T(\Omega_j)\subseteq T(\mathrm{R}_j)$ are equally probable, and that point observations corresponding to time steps outside the route time span $T(\mathrm{R}_j)$ are impossible. The former assumption has no effect on the practical use of $p\big(\Omega_j\,|\,\mathrm{E},\mathrm{R}_j\big)$, since it is used in practice as a function of E and R_j with $T(\Omega_j)$ fixed and known. Both, along with the assumption that the observations $\boldsymbol{\omega}_j^t$ at different time steps are conditionally independent given R_j and depend only on the true position $\mathbf{r}_j^t$ of the object at its corresponding t, lead to

$$p\big(\Omega_j\,|\,\mathrm{E},\mathrm{R}_j\big) \propto \begin{cases} \displaystyle\prod_{t\in T(\Omega_j)} p\big(\boldsymbol{\omega}_j^t\,|\,\mathrm{E},\mathbf{r}_j^t\big), & T(\Omega_j)\subseteq T(\mathrm{R}_j),\\ 0, & T(\Omega_j)\nsubseteq T(\mathrm{R}_j), \end{cases}$$

where the contribution $p\big(\boldsymbol{\omega}_j^t\,|\,\mathrm{E},\mathbf{r}_j^t\big)$ represents the specific observation process at time t.

As for the observation, we assume that each point $\boldsymbol{\omega}_j^t$ is the true position of the corresponding object route R_j with respect to the local (metrically-rectified) ground-plane coordinate system $\Re_c$ associated to the camera plus an additive, normal, homogeneous and isotropic noise. In these conditions we can write

$$p\big(\boldsymbol{\omega}_j^t\,|\,\mathrm{E},\mathbf{r}_j^t\big) \equiv G\big(\boldsymbol{\omega}_j^t;\nu_j^t,\Sigma_\Omega\big), \tag{8}$$

which represents the probability density function (pdf) of the normal distribution $\mathcal{N}(\boldsymbol{\omega}_j^t;\nu_j^t,\Sigma_\Omega)$. The covariance matrix $\Sigma_\Omega=\sigma_\Omega^2\,\mathrm{I}$, assumed constant and isotropic for convenience [36], represents the zero-mean observation noise added to ν_j^t, which represents the route position $\mathbf{r}_j^t$ expressed in the local coordinate system $\Re_c$ as

$$\boldsymbol{\nu}_j^t = m^{-1}\big(\mathbf{r}_j^t;\mathbf{c},\varphi\big) = \big(\mathrm{R}(\varphi)\big)^{-1}\big(\mathbf{r}_j^t-\mathbf{c}\big),$$

where $m^{-1}(\cdot;\mathbf{c},\varphi)$ is the inverse of the linear isometry (1). However, the density $p\big(\boldsymbol{\omega}_j^t\,|\,\mathrm{E},\mathbf{r}_j^t\big)$ is used in (7) as a function of the absolute object position $\mathbf{r}_j^t$, with $\boldsymbol{\omega}_j^t$ assumed fixed. Using the isotropy of the considered distribution and the isometry $m(\cdot;\mathbf{c},\varphi)$ we can rewrite

$$p\big(\boldsymbol{\omega}_j^t\,|\,\mathrm{E},\mathbf{r}_j^t\big) \equiv G\big(\mathbf{r}_j^t;\tilde{\nu}_j^t,\Sigma_\Omega\big),$$

where $\tilde{\nu}_j^t = m(\boldsymbol{\omega}_j^t;\mathbf{c},\varphi)$. We use this equivalent reinterpretation, and not (8).

5.1.2 Route Prior Model

The use of specific information on the scene allows a realistic modeling of object dynamics. The aim of the route prior model is to interpret the characteristics of the map M and translate them into a set of statistical relationships

1. As made clear from the observation model defined in Section 5.1.1, multiples of this framerate could be used without major modifications.

controlling the movement of objects. Our proposed model is divided into two factors concerning, respectively, high-level and low-level features as

$$p(\mathrm{R}_j \,|\, \mathrm{M}) = P(\mathcal{H}_j \,|\, \mathrm{M})\; p\left(T(\mathrm{R}_j), \left(\mathbf{r}_j^t, \mathbf{v}_j^t\right)_{t\in T(\mathrm{R}_j)} \,|\, \mathcal{H}_j, \mathrm{M}\right). \quad (9)$$

The first factor, the probability of a certain high-level path $\mathcal{H}_j$, allows probability definitions based on criteria such as distance, sharpness of turns or privileged areas. We define a prior for $\mathcal{H}_j$ that first selects one entry and one exit node for the route from, respectively, $\mathcal{G}_A$ and $\mathcal{G}_D$, with uniform probability as in principle no information on the origin and destiny of objects will be available, and penalizes longer paths between the two selected $\mathcal{G}$ nodes. Using the representation $\mathcal{H}_j = (\mathfrak{n}_A, \mathcal{H}_j^\circ, \mathfrak{n}_D)$ with $\mathcal{H}_j^\circ$ representing all the intermediate nodes of the path, this can be written as

$$P\big(\mathcal{H}_j \,|\, \mathrm{M}\big) \;=\; P(\mathfrak{n}_A \,|\, \mathrm{M})\, P(\mathfrak{n}_D \,|\, \mathrm{M})\, P\big(\mathcal{H}_j^\circ \,|\, \mathfrak{n}_A, \mathfrak{n}_D, \mathrm{M}\big),$$

where the two first factors are uniform and thus constant for all objects, whereas the latter has been chosen to be proportional to the inverse of a power of the length $l(\mathcal{H}_j)$ of the complete path, penalizing longer paths as:

$$P\big(\mathcal{H}_j^\circ \,|\, \mathfrak{n}_A, \mathfrak{n}_D, \mathrm{M}\big) \propto \big(\, l(\mathcal{H}_j)\big)^{-\alpha}, \quad \alpha > 1.$$

The latter term of the factorization (9) models all the low-level aspects of the route R_j (given $\mathcal{H}_j$), namely its time span $T(\mathrm{R}_j) = (t_j^0, \ldots, t_j^F)$ and the position $\mathbf{r}_j^t$ and velocity $\mathbf{v}_j^t$ of the object at each time step in $T(\mathrm{R}_j)$. As for t_j^0, representing the time when the object first enters the scene, we consider a non-informative (uniform) prior defined on a time range, equal for all the N_Ω objects in the scene, long enough to guarantee that all routes consistent with the tracks Ω_j are included.

We assume that low-level routes of objects can be modeled as a first-order Markov process. This assumption allows the simulation of low-level routes in a simple and fast manner, essential for route marginalization (Appendix A). Moreover, it prevents an artificial explicit modeling of the duration of $T(\mathrm{R}_j)$ since route length can be implicitly considered in the proper Markov model by including, at each time step, a certain disappearance probability $P_D\big(t \,|\, \mathbf{r}_j^t, \mathcal{H}_j, \mathrm{M}\big)$. This probability depends on the distance from the position $\mathbf{r}_j^t$ of the object at each time step to the spatial support $\mathrm{B}_{\mathfrak{n}_D}$ of the exit node $\mathfrak{n}_D$ of the high-level path $\mathcal{H}_j$. This Markov model, depicted graphically in Fig. 4, results in

$$\begin{aligned}
&p\big(T(\mathrm{R}_j), \left(\mathbf{r}_j^t, \mathbf{v}_j^t\right)_{t\in T(\mathrm{R}_j)} \,|\, \mathcal{H}_j, \mathrm{M}\big) \\
&\quad = P_A\big(t_j^0 \,|\, \mathrm{M}\big)\; p_A\left(\mathbf{r}_j^{t_j^0}, \mathbf{v}_j^{t_j^0} \middle| \mathcal{H}_j, \mathrm{M}\right) \\
&\qquad \times \prod_{t=t_j^0}^{t_j^F-1} \left[\big(1 - P_D\big(t \,|\, \mathbf{r}_j^t, \mathcal{H}_j, \mathrm{M}\big)\big) p\big(\mathbf{r}_j^{t+1}, \mathbf{v}_j^{t+1} | \mathbf{r}_j^t, \mathbf{v}_j^t, \mathcal{H}_j, \mathrm{M}\big)\right] \\
&\qquad \times P_D\big(t_j^F \,|\, \mathbf{r}_j^{t_j^F}, \mathcal{H}_j, \mathrm{M}\big),
\end{aligned} \quad (10)$$

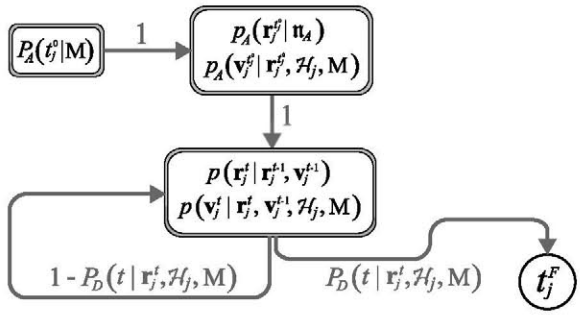

Fig. 4. Prior low-level route distribution, defined as a Markov process.

where we have explicitly indicated that the first time step of the route follows a different model, denoted by $p_A(\cdot)$ from the rest of $T(\mathrm{R}_j)$. The disappearance probability term $P_D\big(t \,|\, \mathbf{r}_j^t, \mathcal{H}_j, \mathrm{M}\big)$ is a binary function such that objects always continue their way whenever $\mathbf{r}_j^t$ has not reached $\mathrm{B}_{\mathfrak{n}_D}$ (which is equivalent to disappearance probability identically zero) and always disappear immediately when they reach $\mathrm{B}_{\mathfrak{n}_D}$ (disappearance probability identically one).

As for the Markovian term for position and velocity time evolution, we use the decomposition

$$\begin{aligned}
&p\big(\mathbf{r}_j^t, \mathbf{v}_j^t \,|\, \mathbf{r}_j^{t-1}, \mathbf{v}_j^{t-1}, \mathcal{H}_j, \mathrm{M}\big) \\
&\quad = p\big(\mathbf{r}_j^t \,|\, \mathbf{r}_j^{t-1}, \mathbf{v}_j^{t-1}\big) p\big(\mathbf{v}_j^t \,|\, \mathbf{r}_j^t, \mathbf{v}_j^{t-1}, \mathbf{r}_j^{t-1}, \mathcal{H}_j, \mathrm{M}\big),
\end{aligned}$$

which allows the independent definition of position and velocity evolution models (dependencies are explicitly indicated). The first term is the distribution of the position $\mathbf{r}_j^t$ and reflects the assumption that the velocity $\mathbf{v}_j^{t-1}$ at time $t-1$ was chosen so as to cause a reasonable object location at t, and that no major deviations from the resulting linear evolution are experienced. For this reason, we set

$$p\big(\mathbf{r}_j^t \,|\, \mathbf{r}_j^{t-1}, \mathbf{v}_j^{t-1}\big) \equiv G\big(\mathbf{r}_j^t;\, \boldsymbol{\eta}_j^t, \Sigma_\mathrm{r}\big), \quad (11)$$

where $\Sigma_\mathrm{r} = \sigma_\mathrm{r}^2\, \mathrm{I}$ represents the covariance matrix of the isotropic zero-mean normal prediction noise added to the expected position $\boldsymbol{\eta}_j^t = \mathbf{r}_j^{t-1} + \mathbf{v}_j^{t-1}$.

As said above, $\mathbf{v}_j^t$ is chosen so as to generate a satisfactory position $\mathbf{r}_j^{t+1}$: thus, $p(\mathbf{v}_j^t \,|\, \mathbf{r}_j^t, \mathbf{v}_j^{t-1}, \mathbf{r}_j^{t-1}, \mathcal{H}_j, \mathrm{M})$ is responsible for reflecting the main characteristics of the trajectories of the specific moving objects. For this reason, it should be defined taking into account the specific environment (indoor/outdoor, urban) and moving object type (vehicles, pedestrians). To illustrate the creation of $p(\mathbf{v}_j^t \,|\, \mathbf{r}_j^t, \mathbf{v}_j^{t-1}, \mathbf{r}_j^{t-1}, \mathcal{H}_j, \mathrm{M})$, we use here an urban traffic model. However, other types of environments and moving objects would be equally compatible with the presented framework.

Our definitions for urban traffic dynamics stem from the following assumptions: (i) local behavior of vehicles with respect to the local characteristics of their trajectories varies slowly; (ii) vehicles tend to move along the tangential direction of their high-level path $\mathcal{H}_j$ and to a certain average speed V_{AVG} which depends on the scene itself; and (iii) vehicles keep their position within the passable regions composing the high-level path $\mathcal{H}_j$, tending to leave a certain distance W between their expected

position $\mathbf{r}_j^{t+1}$ and the limit of $\mathcal{H}_j$. We use two auxiliary structures to define a practical $p(\mathbf{v}_j^t \,|\, \mathbf{r}_j^t, \mathbf{v}_j^{t-1}, \mathbf{r}_j^{t-1}, \mathcal{H}_j, \mathrm{M})$ fulfilling these properties: the signed Euclidean distance $d_{\mathcal{H}_j}(\mathbf{r})$ to the limit of the binary mask $\mathrm{B}_{\mathcal{H}_j}$ of $\mathcal{H}_j$ (with negative values inside $\mathcal{H}_j$ passable regions, and positive outside) and the unitary tangential vector field $\tau_{\mathcal{H}_j}(\mathbf{r})$, orientated towards the direction of vehicles along the path $\mathcal{H}_j$. The former is used for influencing or "correcting" vehicle positions near the limits of the underlying high-level path. The latter allows the definition of a "moving" local coordinate system indicating, at each point, the tangent and normal directions of the given path, making thus possible to propagate the local behavior of objects along their trajectory. Both $d_{\mathcal{H}_j}(\mathbf{r})$ and $\tau_{\mathcal{H}_j}(\mathbf{r})$, and all the structures derived from them, will be calculated on a discrete spatial grid through simple operations performed on the binary mask $\mathrm{B}_{\mathcal{H}_j}$ and evaluated later on any continuous position using interpolation. Using these structures, we define

$$p\big(\mathbf{v}_j^t \,|\, \mathbf{r}_j^t, \mathbf{v}_j^{t-1}, \mathbf{r}_j^{t-1}, \mathcal{H}_j, \mathrm{M}\big) \equiv G\big(\mathbf{v}_j^t;\, \boldsymbol{\mu}_j^t, \Sigma_{\mathbf{v}}\big),$$

where $\Sigma_{\mathbf{v}} = \sigma_{\mathbf{v}}^2\, \mathrm{I}$ is the covariance matrix of the isotropic zero-mean normal prediction noise added to an expected velocity defined as the sum $\boldsymbol{\mu}_j^t = \overline{\overline{\mathbf{v}}}_j^{t-1} + \mathbf{v}_c(\mathbf{r}_j^t, \overline{\overline{\mathbf{v}}}_j^{t-1})$, where $\overline{\overline{\mathbf{v}}}_j^{t-1}$ represents the adaptation of the velocity $\mathbf{v}_j^{t-1}$ of the object at $\mathbf{r}_j^{t-1}$ to the local characteristics of its path at $\mathbf{r}_j^t$, and where $\mathbf{v}_c$ represents a correction function aimed at keeping object position within the path.

As for the adapted velocity term $\overline{\overline{\mathbf{v}}}_j^{t-1}$, we assume that the tangent and normal components of $\mathbf{v}_j^{t-1}$ are kept and adapted individually as

$$\begin{aligned}\overline{\overline{\mathbf{v}}}_j^{t-1} = {} & g_\tau\big(\mathbf{v}_j^{t-1} \cdot \tau_{\mathcal{H}_j}(\mathbf{r}_j^{t-1})\big)\, \tau_{\mathcal{H}_j}(\mathbf{r}_j^t) \\ & + g_n\big(\mathbf{v}_j^{t-1} \cdot \mathbf{n}_{\mathcal{H}_j}(\mathbf{r}_j^{t-1})\big)\, \mathbf{n}_{\mathcal{H}_j}(\mathbf{r}_j^t),\end{aligned}$$

where $\mathbf{n}_{\mathcal{H}_j}(\mathbf{r})$ represents the normal vector that forms, along with $\tau_{\mathcal{H}_j}(\mathbf{r})$, a positive reference system at $\mathbf{r}$. To encourage routes with tangential speed close to V_{AVG}, the function $g_\tau(\cdot)$ controlling the scalar evolution of the tangential velocity component is defined as

$$g_\tau(V) = g_1 V + g_0, \quad \text{with} \begin{cases} 0 < g_0 < V_{\mathrm{AVG}}, \\ g_1 = 1 - g_0 / V_{\mathrm{AVG}}, \end{cases}$$

contractive with fixed point at V_{AVG}. Analogously, to encourage objects to move exclusively along tangential trajectories, the function $g_n(\cdot)$ controlling the scalar evolution of the normal component is defined contractive and antisymmetric (thus with fixed point at the origin) according to

$$g_n(V) = g_2 V\,,\ 0 < g_2 < 1.$$

The distance correction term $\mathbf{v}_c(\mathbf{r}_j^t, \overline{\overline{\mathbf{v}}}_j^{t-1})$ aims to compensate the expected position $(\mathbf{r}_j^t + \overline{\overline{\mathbf{v}}}_j^{t-1})$ so as to keep the resulting distance $d_{\mathcal{H}_j}$ to the border of the path below zero (i.e., inside its mask $\mathrm{B}_{\mathcal{H}_j}$), leaving velocity unaltered when the expected position has signed distance lower than $-W$ for a certain positive W. Our proposed correction factor is of the form $\mathbf{v}_c(\mathbf{r}_j^t, \overline{\overline{\mathbf{v}}}_j^{t-1}) = v_c(\mathbf{r}_j^t, \overline{\overline{\mathbf{v}}}_j^{t-1})\, \nabla d_{\mathcal{H}_j}(\mathbf{r})$ for a certain scalar function $v_c(\cdot)$, whose definition stems from the approximation

$$d_{\mathcal{H}_j}\big(\mathbf{r}_j^t + \overline{\overline{\mathbf{v}}}_j^{t-1} + \mathbf{v}_c\big) \approx d_{\mathcal{H}_j}\big(\mathbf{r}_j^t\big) + \overline{\overline{\mathbf{v}}}_j^{t-1} \cdot \nabla d_{\mathcal{H}_j}(\mathbf{r}) + v_c,$$

where we have used that $\|\nabla d_{\mathcal{H}_j}\| = 1$ for the Euclidean norm in the regions of interest (note that it represents the spatial rate of change of the spatial distance). The scalar correction function $v_c(d)$ is then defined as

$$v_c(d) = \begin{cases} 0, & d \le (-W), \\ -(d + W \exp\{-(1 + d/W)\}), & d \ge (-W), \end{cases}$$

which satisfies the above requirements and which obviously reduces the distance of all object positions such that $-W < d_{\mathcal{H}_j} < \infty$. The velocity evolutionary model for vehicles is illustrated in Fig. 5.

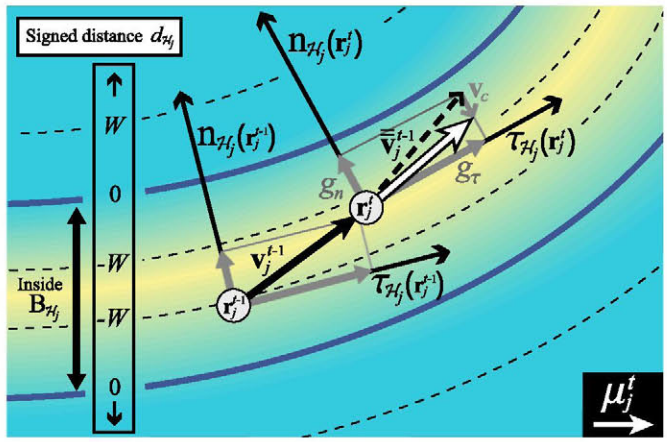

Fig. 5. Schematic description of the velocity evolution of objects: computation of $\boldsymbol{\mu}_j^t$. Background gradual colouring indicates $d_{\mathcal{H}_j}(\mathbf{r})$, from yellow (negative) to blue (positive).

The prior distribution for the position and velocity of the vehicles at their initial time step t_j^0 is defined as

$$p_A\Big(\mathbf{r}_j^{t_j^0}, \mathbf{v}_j^{t_j^0} \,|\, \mathcal{H}_j, \mathrm{M}\Big) = p_A\Big(\mathbf{r}_j^{t_j^0} \,|\, \mathfrak{n}_A\Big)\; p_A\Big(\mathbf{v}_j^{t_j^0} \,|\, \mathbf{r}_j^{t_j^0}, \mathcal{H}_j, \mathrm{M}\Big).$$

Both initial position and velocity are considered isotropic and normally distributed, and centered respectively at the expected entry point $\mathbf{b}_{\mathfrak{n}_A}$ of the entry node $\mathfrak{n}_A$ and the expected tangential velocity $\boldsymbol{\mu}_j^{t_j^0} = V_{\mathrm{AVG}}\, \tau_{\mathcal{H}_j}(\mathbf{r}_j^{t_j^0})$.

Fig. 6 displays different routes simulated using the proposed dynamic model for a certain high-level path $\mathcal{H}_j$, showing that our proposal is able to capture the restrictions imposed by the environment.

5.2 Numerical Evaluation of the Posterior

The posterior pdf $p(\mathrm{E} \,|\, \Omega, \mathrm{M})$ is defined as the marginalization of $p(\mathrm{E}, \mathrm{R} \,|\, \Omega, \mathrm{M})$, factorized as (7), over R. This process can be divided into two steps: first, marginalization of $p(\mathrm{E}, \mathrm{R} \,|\, \Omega, \mathrm{M})$ over the low-level details of all R_j ($j \in \mathcal{J}$), which can be written as

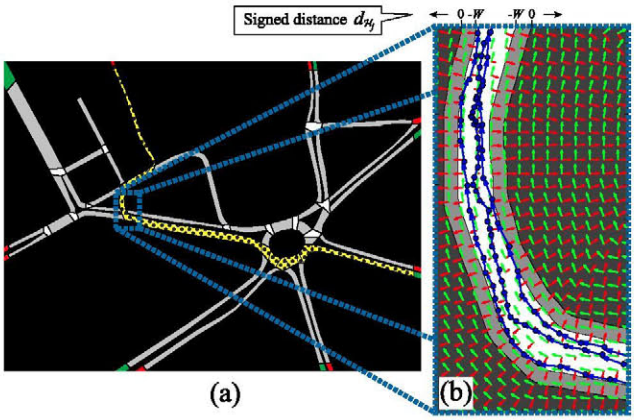

Fig. 6. Example of route simulation with the proposed dynamic model. (a) First, high-level path $\mathcal{H}_j$ sampling (dotted). (b) Low-level route generation using the fields $\tau_{\mathcal{H}_j}(\mathbf{r})$ and $\nabla d_{\mathcal{H}_j}$ (green and red) and partial magnification of three simulated routes.

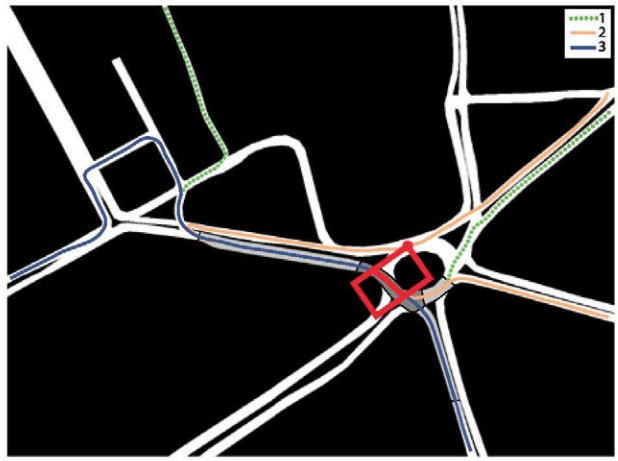

Fig. 7. High-level path grouping performed at the principal module. Three paths enclosing the true absolute 2D track and thus sharing the same central nodes (depicted as low-level routes). Unlike path 3, paths 1 and 2 belong to the same class (same nodes around the track).

$$
\begin{aligned}
&p(\mathrm{E}, (\mathcal{H}_j)_{j\in\mathcal{J}} \,|\, (\Omega_j)_{j\in\mathcal{J}}, \mathrm{M}) \\
&\propto \prod_{j\in\mathcal{J}} \left[\sum_{\forall T(\mathrm{R}_j)} \int_{(\mathbf{r}_j^t, \mathbf{v}_j^t)_{t\in T(\mathrm{R}_j)}} p\big(\Omega_j \,|\, \mathrm{E}, \mathrm{R}_j\big)\, p\big(\mathrm{R}_j \,|\, \mathrm{M}\big)\, \mathrm{d}\big(\mathbf{r}_j^t, \mathbf{v}_j^t\big)_{t\in T(\mathrm{R}_j)} \right] \\
&= \prod_{j\in\mathcal{J}} \left[\sum_{\forall T(\mathrm{R}_j)} \int_{(\mathbf{r}_j^t, \mathbf{v}_j^t)_{t\in T(\mathrm{R}_j)}} p\big(\Omega_j \,|\, \mathrm{E}, \mathrm{R}_j\big)\; P\big(\mathcal{H}_j \,|\, \mathrm{M}\big) \right. \\
&\qquad \left. p\big(T(\mathrm{R}_j), \big(\mathbf{r}_j^t, \mathbf{v}_j^t\big)_{t\in T(\mathrm{R}_j)} \,|\, \mathcal{H}_j, \mathrm{M}\big) \mathrm{d}\big(\mathbf{r}_j^t, \mathbf{v}_j^t\big)_{t\in T(\mathrm{R}_j)} \right] \\
&= \prod_{j\in\mathcal{J}} [\, P(\mathcal{H}_j \,|\, \mathrm{M})\; h(\mathrm{E}, \mathcal{H}_j; \Omega_j)],
\end{aligned}
\tag{12}
$$

where $\mathrm{d}(\mathbf{r}_j^t, \mathbf{v}_j^t)_{t\in T(\mathrm{R}_j)}$ indicates that integration is carried out over all positions and velocities in $T(\mathrm{R}_j)$ and with

$$
\begin{aligned}
h(\mathrm{E}, \mathcal{H}_j; \Omega_j) = \sum_{\forall T(\mathrm{R}_j)} \int_{(\mathbf{r}_j^t, \mathbf{v}_j^t)_{t\in T(\mathrm{R}_j)}} & p\big(\Omega_j \,|\, \mathrm{E}, \mathrm{R}_j\big) \\
& p\big(T(\mathrm{R}_j), \big(\mathbf{r}_j^t, \mathbf{v}_j^t\big)_{t\in T(\mathrm{R}_j)} \,|\, \mathcal{H}_j, \mathrm{M}\big) \mathrm{d}\big(\mathbf{r}_j^t, \mathbf{v}_j^t\big)_{t\in T(\mathrm{R}_j)},
\end{aligned}
\tag{13}
$$

and second, marginalization of (12) over the discrete subspace of possible high-level routes across the map to isolate completely the camera position E, obtaining

$$
\begin{aligned}
p(\mathrm{E} \,|\, (\Omega_j)_{j\in\mathcal{J}}, \mathrm{M}) &= \sum_{\forall\{(\mathcal{H}_j)_{j\in\mathcal{J}}\}} p(\mathrm{E}, (\mathcal{H}_j)_{j\in\mathcal{J}} \,|\, (\Omega_j)_{j\in\mathcal{J}}, \mathrm{M}) \\
&\propto \prod_{j\in\mathcal{J}} \sum_{\forall\mathcal{H}_j} [P\big(\mathcal{H}_j \,|\, \mathrm{M}\big)\, h(\mathrm{E}, \mathcal{H}_j; \Omega_j)].
\end{aligned}
\tag{14}
$$

The evaluation of $p\big(\mathrm{E} \,|\, \Omega, \mathrm{M}\big)$ is based on the estimation of the sum/integral unit $h(\mathrm{E}, \mathcal{H}_j; \Omega_j)$ for each high-level path $\mathcal{H}_j$ of each object. The MCMC sampling process from $p\big(\mathrm{E} \,|\, \Omega, \mathrm{M}\big)$ (Section 5.3) requires the frequent evaluation of (14): for this reason, we have designed an efficient algorithm for estimating $h(\mathrm{E}, \mathcal{H}_j; \Omega_j)$, based on sequential Monte Carlo methods and importance sampling as detailed in Appendix A.

We also apply high-level path grouping to improve the efficiency of the marginalization process. This simplification is rooted in the fact that the calculation of $h(\mathrm{E}, \mathcal{H}_j; \Omega_j)$ involves only the observations $\boldsymbol{\omega}_j^t$, expressed as $m(\boldsymbol{\omega}_j^t; \mathrm{E})$ with respect to the absolute frame of reference $\Re_w$ (where $m(\cdot; \mathrm{E})$ is the isometry (1)): thus, high-level paths differing only in aspects that do not affect the low-level route distribution in the areas around $m(\boldsymbol{\omega}_j^t; \mathrm{E})$ must provide similar $h(\mathrm{E}, \mathcal{H}_j; \Omega_j)$. Low-level dynamics defined in Section 5.1.2 are determined by the fields $d_{\mathcal{H}_j}(\mathbf{r})$ and $\tau_{\mathcal{H}_j}(\mathbf{r})$, calculated using local characteristics of the mask $\mathrm{B}_{\mathcal{H}_j}$ of the high-level path $\mathcal{H}_j$: therefore, all paths $\mathcal{H}_j$ having an identical $\mathrm{B}_{\mathcal{H}_j}$ "before" (according to the direction of $\mathcal{H}_j$) and in the area where the observations $m(\boldsymbol{\omega}_j^t; \mathrm{E})$ lie must yield the same $h(\mathrm{E}, \mathcal{H}_j; \Omega_j)$. For this reason, all those $\mathcal{H}_j$ whose nodes coincide until the position of the last of the point observations $m(\boldsymbol{\omega}_j^t; \mathrm{E})$ are grouped, and $h(\mathrm{E}, \mathcal{H}_j; \Omega_j)$ is only evaluated once for them all.

Additionally, although it is not exact, we can follow a similar reasoning for those nodes composing the high-level path $\mathcal{H}_j$ "before" the observations $m(\boldsymbol{\omega}_j^t; \mathrm{E})$, $t \in T(\Omega_j)$, and assume that the distribution of route low-level characteristics "forgets" its details after a certain distance. We perform thus a previous analysis for grouping high-level paths into a reduced set of classes, characterized by sharing the same preceding N_{pre} and posterior N_{post} nodes around the observations of the corresponding object. Integral calculations are then performed only for one representative of each class. This simplification (Fig. 7) reduces drastically the number of integrals and therefore the computational cost associated to the evaluation of $p(\mathrm{E} \,|\, \Omega, \mathrm{M})$. Fig. 8 shows that the differences between the integrals of different elements belonging to the same class ($N_{\mathrm{pre}} = N_{\mathrm{post}} = 2$) are much lower than the variance of the integration algorithm itself, which justifies the grouping process.

5.3 MCMC Sampling and Move Definition

The proposed sampling process based on the Metropolis-Hastings algorithm [37] considers two moves: move 1 (m_1), the modification of a previously accepted camera hypothesis; and move 2 (m_2), the proposal of a new camera hypothesis from the data-driven proposal $q\big(\mathrm{E} \,|\, \mathrm{M}\big)$ estimated in the preprocessing module, needed to reach distant regions of

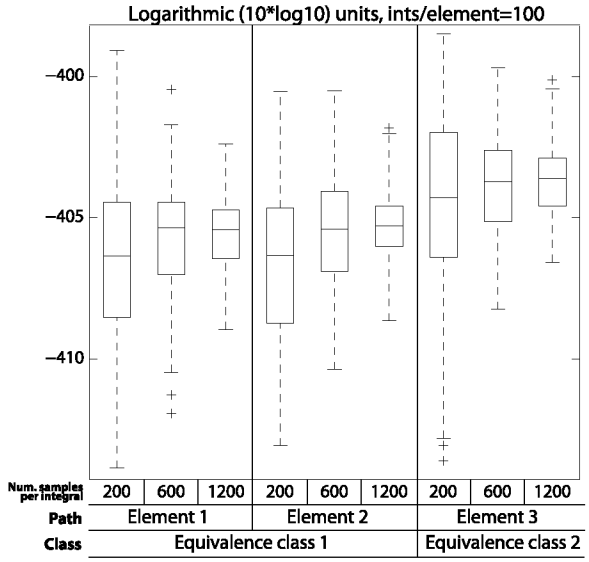

Fig. 8. Evaluation of the integral $h(\mathrm{E},\mathcal{H}_j;\Omega_j)$ for the three paths in Fig. 7 (100 different runs per path and number of samples, probability expressed in logarithmic units). Relevant statistical indicators showed as a box plot.

the solution subspace. The moves will be chosen with probability $q(m_1)$ and $q(m_2)$, respectively, with $q(m_1)+q(m_2)=1$.

Move 1 uses a proposal kernel $q(\mathrm{E}\,|\,\check{\mathrm{E}})$ normal and centered at the previously accepted sample $\check{\mathrm{E}}$ to draw a new hypothesis $\hat{\mathrm{E}}$. Thus, $q(\hat{\mathrm{E}}\,|\,\check{\mathrm{E}}) = q(\check{\mathrm{E}}\,|\,\hat{\mathrm{E}})$, and the resulting acceptance ratio for a new hypothesis $\hat{\mathrm{E}}$ is $\overline{\alpha}_1 = \min\{1,\alpha_1\}$, where

$$\begin{aligned}\alpha_1 &= \frac{p(\hat{\mathrm{E}}\,|\,\Omega,\mathrm{M})}{p(\check{\mathrm{E}}\,|\,\Omega,\mathrm{M})}\frac{q(m_1)\;q(\check{\mathrm{E}}\,|\,\hat{\mathrm{E}})}{q(m_1)\;q(\hat{\mathrm{E}}\,|\,\check{\mathrm{E}})}\\ &= \frac{\prod_{j\in\mathcal{J}}\left[\sum_{\forall\mathcal{H}'_j} P(\mathcal{H}'_j\,|\,\mathrm{M})h(\hat{\mathrm{E}},\mathcal{H}'_j;\Omega_j)\right]}{\prod_{j\in\mathcal{J}}\left[\sum_{\forall\mathcal{H}_j} P(\mathcal{H}_j\,|\,\mathrm{M})h(\check{\mathrm{E}},\mathcal{H}_j;\Omega_j)\right]},\end{aligned} \tag{15}$$

and where $h(\mathrm{E},\mathcal{H}_j;\Omega_j)$ represents the low-level route integration unit discussed in Appendix A. In move 2, where the generation of new hypotheses $\hat{\mathrm{E}}$ is directed by the kernel-based proposal $\tilde{q}(\mathrm{E}|\mathrm{M})$, acceptance will be driven by the ratio $\overline{\alpha}_2 = \min\{1,\alpha_2\}$, where

$$\alpha_2 = \frac{p(\hat{\mathrm{E}}\,|\,\Omega,\mathrm{M})}{p(\check{\mathrm{E}}\,|\,\Omega,\mathrm{M})}\frac{q(m_2)\;q(\check{\mathrm{E}}\,|\,\mathrm{M})}{q(m_2)\;q(\hat{\mathrm{E}}\,|\,\mathrm{M})} = \alpha_1\frac{q(\check{\mathrm{E}}\,|\,\mathrm{M})}{q(\hat{\mathrm{E}}\,|\,\mathrm{M})},$$

where, unlike (15), proposal densities do not cancel mutually and must thus be explicitly calculated.

Accepted sample rate strongly depends on how well the seeds generated in the purely geometric analysis of the preprocessing module fit the dynamics of the scene, since certain observations can geometrically fit certain areas of the map but have negligible probability once object dynamics are considered. To prevent the unnecessary evaluation of clearly erroneous camera hypotheses, we eliminate every checked proposal seed that has proved against the dynamics of the map. This simple action eliminates most erroneous seeds during the burn-in phase of MCMC, and improves the acceptance rate of subsequent iterations.

5.4 Hypothesis Selection and Ambiguity Analysis

The marginal posterior $p(\mathrm{E}\,|\,\Omega,\mathrm{M})$ is composed of an indeterminate high number of probability modes sparsely distributed over the solution space. We summarize the main characteristics of this distribution, that is, its main modes and their relative importance, in a reduced set of distinct weighed camera hypotheses $\{\overline{\mathrm{E}}^{(k)},\overline{w}^{(k)}\}_{k=1}^{K}$ using an adaptation of the K-adventurers algorithm [31] for multiple hypothesis preservation. For the sake of readability, we will denote $p(\mathrm{E}\,|\,\Omega,\mathrm{M})$ by $p(\mathrm{E})$.

Let $p(\mathrm{E})$ be a certain target pdf, and let

$$\overline{p}(\mathrm{E}) \equiv \sum_{k=1}^{K}\overline{w}^{(k)}\,\overline{G}\big(\mathrm{E}-\overline{\mathrm{E}}^{(k)}\big) \tag{16}$$

be a kernel-based approximation with exactly K modes, where $\overline{G}(\mathrm{E})$ represents a kernel profile centered at the origin and with fixed scale (adapted to the expected size of the searched modes). We aim to find the set $\{\overline{\mathrm{E}}^{(k)},\overline{w}^{(k)}\}_{k=1}^{K}$ of weighted solutions that best approximates $p(\mathrm{E})$ in terms of the Kullback-Leibler divergence

$$D_{KL}\big(p\|\overline{p}\,\big) = \int p(\mathrm{E})\,\ln\!\left(\frac{p(\mathrm{E})}{\overline{p}(\mathrm{E})}\right)\mathrm{dE} = -H(p) + H(p,\overline{p}),$$

where $H(p)$ and $H(p,\overline{p})$ are, respectively, the entropy of $p(\mathrm{E})$ and the cross entropy of $p(\mathrm{E})$ and $\overline{p}(\mathrm{E})$. Since $H(p)$ is constant for all possible solution sets, this problem is equivalent to the minimization of

$$H(p,\overline{p}) = -\mathbb{E}_{p(\mathrm{E})}\left\{\ln\left[\sum_{k=1}^{K}\overline{w}^{(k)}\,\overline{G}\big(\mathrm{E}-\overline{\mathrm{E}}^{(k)}\big)\right]\right\},$$

where $\mathbb{E}_{p(\mathrm{E})}[\,\cdot\,]$ is the expected value with respect to the reference probability density function $p(\mathrm{E})$. $H(p,\overline{p})$ can be inferred using the RMSE estimator for the mean as

$$H(p,\overline{p}) \approx -\frac{1}{U}\sum_{u=1}^{U}\ln\left[\sum_{k=1}^{K}\overline{w}^{(k)}\,\overline{G}\Big(\overline{\overline{\mathrm{E}}}^{(u)}-\overline{\mathrm{E}}^{(k)}\Big)\right], \tag{17}$$

where $\overline{\overline{\mathrm{E}}}^{(u)} \sim p(\mathrm{E})$. In our case, the U samples are the result of the MCMC step discussed in Section 5.3.

The best summarizing set $\{\overline{\mathrm{E}}^{(k)},\overline{w}^{(k)}\}_{k=1}^{K}$ is searched amongst the samples $\{\overline{\overline{\mathrm{E}}}^{(u)}\}_{u=1}^{U}$. Subsets of K samples are randomly chosen and analyzed, and the best in terms of $H(p,\overline{p})$, along with its corresponding weights $\{\overline{w}^{(k)}\}_{k=1}^{K}$, is finally retained. The weights $\{\overline{w}^{(k)}\}_{k=1}^{K}$ for each test set $\{\overline{\mathrm{E}}^{(k)}\}_{k=1}^{K}$ are suboptimally estimated using Lagrange optimization on a conveniently simplified version of (17), which stems from the assumption that the K modes are separated enough with respect to the scale of the kernels $\overline{G}(\mathrm{E})$. In these conditions, the area where $\overline{p}(\mathrm{E})$ is significant can be fragmented into subregions $\{Z^{(k)}\}_{k=1}^{K}$ where $\overline{\mathrm{E}}^{(k)} \in Z^{(k)}$ and the contribution of the corresponding mode $\overline{\mathrm{E}}^{(k)}$ is totally dominant. Thus, the exact expression

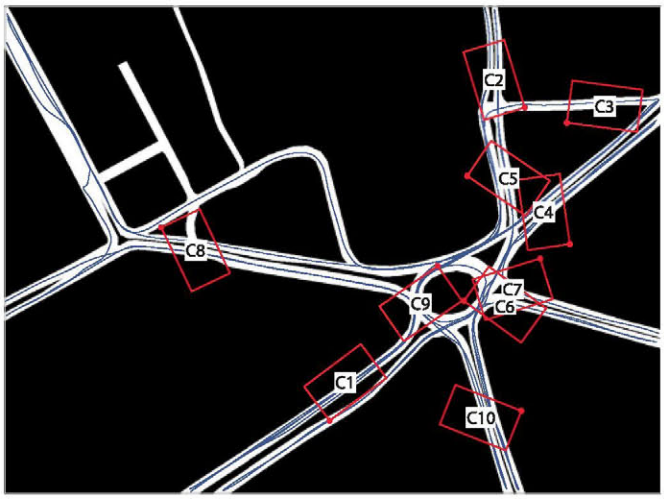

Fig. 9. Synthetic database: object routes (blue), fields of view of the cameras (red).

(16) for $\overline{p}(\mathrm{E})$ can be approximated by

$$\tilde{p}(\mathrm{E}) = \sum_{k=1}^{K} \left[\overline{w}^{(k)}\, \overline{G}\left(\mathrm{E} - \overline{\mathrm{E}}^{(k)}\right) I_{Z^{(k)}}(\mathrm{E})\right],$$

where $I_{Z^{(k)}}(\cdot)$ is the indicator function. This region-based approximation allows us to write the logarithm of the sum as a sum of region-related logarithms, which transforms (17) into

$$\begin{aligned}\tilde{H}(p,\overline{p}) &= -\frac{1}{U}\sum_{u=1}^{U} \ln\left[\sum_{k=1}^{K} \overline{w}^{(k)} \overline{G}\left(\overline{\overline{\mathrm{E}}}^{(u)} - \overline{\mathrm{E}}^{(k)}\right) I_{Z^{(k)}}\left(\overline{\overline{\mathrm{E}}}^{(u)}\right)\right] \\ &= -\frac{1}{U}\sum_{u=1}^{U}\sum_{k=1}^{K}\left\{\ln\left[\overline{w}^{(k)} \overline{G}\left(\overline{\overline{\mathrm{E}}}^{(u)} - \overline{\mathrm{E}}^{(k)}\right)\right] I_{Z^{(k)}}\left(\overline{\overline{\mathrm{E}}}^{(u)}\right)\right\}.\end{aligned} \tag{18}$$

This last expression is suitable for the method of Lagrange multipliers, applied to find the set $\{\overline{w}^{(k)}\}_{k=1}^{K}$ that minimize $\tilde{H}(p,\overline{p})$ subject to the restriction $\sum_{k=1}^{K} \overline{w}^{(k)} = 1$ (note that the additional restriction $\overline{w}^{(k)} \geq 0$ is in principle required, but has no effect in this case as shown by the result). The construction of the Lagrange function

$$\Lambda(\mathbf{w},\lambda) = \tilde{H}(p,\overline{p}) + \lambda\left(\sum_{k=1}^{K} \overline{w}^{(k)} - 1\right)$$

and the resolution of its resulting system of equations $\nabla\Lambda(\mathbf{w},\lambda) = 0$ yield the final weights

$$\overline{w}^{(k)} = \frac{1}{U}\sum_{u=1}^{U} I_{Z^{(k)}}\left(\overline{\overline{\mathrm{E}}}^{(u)}\right), \forall k \in \{1,\ldots,K\}.$$

Although the preceding formulae are totally general, we assume a kernel profile $\overline{G}(\mathrm{E})$ composed of independent Gaussian (pseudo-Gaussian in the angular component) functions for each dimension of the solution space.

The inference of the number K of significantly distinct modes is important itself to understand the ambiguity of the given setting. Our proposal for its automatical inference iteratively increases in one unit the number $\tilde{K}$ of assumed modes, whose corresponding (sub)optimal KL-divergence is estimated and compared to the best divergence with $\tilde{K}-1$ modes. The algorithm ends when the gain obtained by adding a new mode is below a certain percentage of the best divergence hitherto.

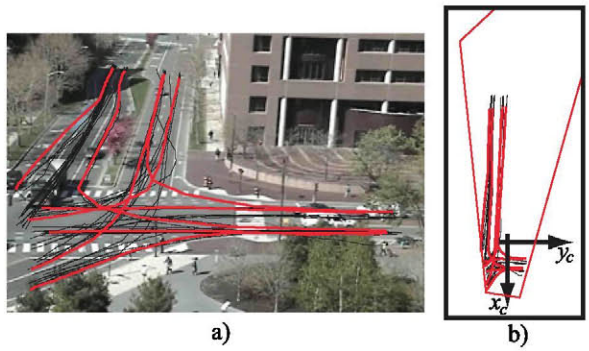

Fig. 10. Real vehicle data set, camera 1: a) View. b) Local ground-plane frame of reference, once camera-to-ground homography has been applied. Black lines: 2D tracks for all objects included in the data set. Red lines: the eight representative tracks retained after trajectory clustering [33].

6 Experimental Validation

6.1 Experimental Setup

The proposed framework and the dynamic model discussed in Section 5.1.2 are tested using both synthetic data and a real database. All the results have been generated using $g_0 = V_{\mathrm{AVG}}/20$ and $g_2 = 3/4$ in the dynamic model definition (values within a reasonable range around these show no major deviations), and a low number of orientations (≈ 40) for integral evaluation in the preprocessing module.

The synthetic database consists of a real road scene for map definition and semi-automatically generated trajectories. The map covers an urban extent of 500×400 m, and its associated binary masks have a resolution of 0.33 m/pixel (Fig. 9). Object average speed was 35 km/h, test routes were generated at a framerate of 5 fps, and the FoV covered by the camera is rectangular (equivalent to a vertically-oriented camera) and of size 50×35 m. Ten different cameras have been randomly generated for testing, each one with a number of observed 2D tracks between 4 and 10.

The real vehicle database comprises four traffic cameras (referred to as MIT-1 to MIT-4) imaging entries to and exits from parking lots. MIT-1 and its associated tracks have been adapted from the MIT Traffic Data Set [38], while MIT-2, 3 and 4 are from the MIT Trajectory Data Set [39]. The camera views are related to the scene ground plane using a metric rectification homography calculated manually using the DLT algorithm [40], and the resulting "corrected" views are considered the local frame of reference of each camera (Fig. 10). The map covers $1{,}200 \times 850$ m, and its associated masks have a resolution of 0.4 m/pixel. Object average speed is 35 km/h, and the framerate of the observed 2D tracks is 4.83 fps (1/6 of the original framerate of MIT-1, 29 fps). MIT-1 contains 93 real vehicle trajectories, but only eight different representative tracks were retained after the trajectory clustering procedure [33] (Fig. 10). MIT-2, 3 and 4

Fig. 11. Camera position hypotheses (green) sampled from the proposal distribution $q(\mathrm{E} \mid (\Omega_j)_{j \in \mathcal{J}}, \mathrm{M})$ estimated at the preprocessing module for camera C1. Input data contaminated with observational noise of increasing power, and similar obtained results. True camera location (red) and KL-representatives (blue) have been included for clarity.

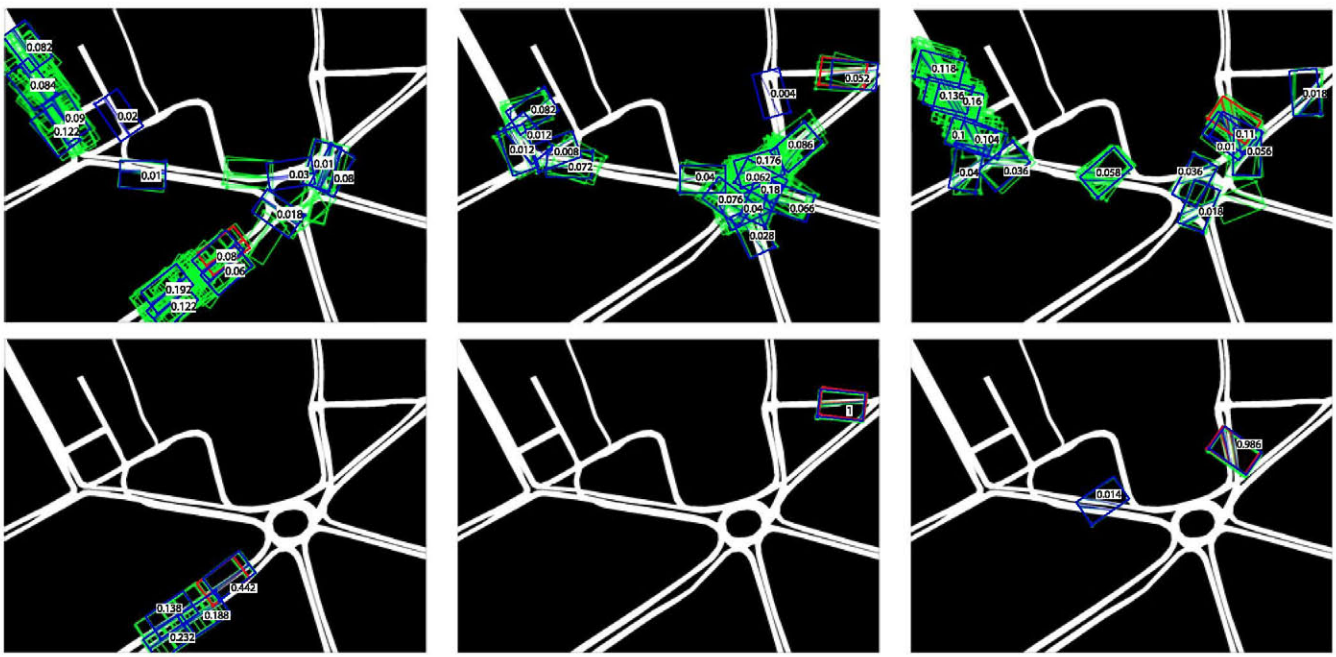

Fig. 12. Preprocessing module proposals (first row) versus principal module results (second row). In green, camera samples generated in each case. True camera location (red) and KL-representatives (blue), with relative weights superimposed, have been included to indicate the estimated uncertainty.

contain, respectively, 4,025, 4,096 and 5,753 tracks, approximately half of them corresponding to pedestrians: after the clustering procedure, 7, 10 and 9 representative vehicle tracks were retained for each camera.

Additional details on the real vehicle databases, especially on the characteristics of the considered camera views and on the urban structures observed from them, can be found in our website.[2] The site contains also more detailed descriptions for the synthetic settings, which helps to better understand the results discussed in the next section.

6.2 Performance Evaluation

First, we evaluate the preprocessing module and the resulting camera position proposal $q(\mathrm{E} \mid \Omega, \mathrm{M})$. Fig. 11 presents results corresponding to the same camera setting but with 2D tracks contaminated with different observational noise power. All cases calculate the proposal using 1,000 seeds, and display 100 randomly-generated camera hypotheses showing that the proposals are resilient to noise. Note that the true camera location is included between those that can be proposed by $q(\mathrm{E} \mid \Omega, \mathrm{M})$.

The validity of the integration process used extensively in the principal module and based on high-level class grouping and iterative MCMC evaluation has been demonstrated in Section 5.3. The K-Adventurer-based analysis has been tested individually, and the corresponding experiments have shown that thresholds between 5 and 10 percent provide satisfactory estimates for K in most situations. Presented results have been obtained using 5 percent.

As for the overall performance of the principal module, three illustrative results of its effect are included in Fig. 12 showing the difference between the samples proposed before and after object dynamics consideration. In all cases, the length of the burn-in phase of the MCMC process have been chosen comparable to the number of seeds used by the proposal $q(\mathrm{E} \mid \Omega, \mathrm{M})$ to eliminate as many wrong seeds as possible before the proper MCMC sampling. We

2. http://www.gti.ssr.upm.es/data/CamLocalizationMaps.

TABLE 1
Absolute Error of Camera Estimations Performed at the Principal Module versus Baseline Method
(∗: Straight Zone, Linear Drift, ∗∗: Ambiguous Setting)

Cam	Proposed approach (principal module)			Baseline method	
	Ambiguity detected	Distance error (m)	Angular error (rad)	Distance error (m)	Angular error (rad)
* C1	YES	58.338	0.0363	77.229	0.0421
C2	no	0.731	0.0091	297.029	0.5608
C3	no	0.641	0.0107	348.045	0.6789
**C4	YES	134.088	2.2169	322.226	1.2961
C5	no	3.214	0.0331	304.434	0.2474
C6	no	1.345	0.0120	2.119	0.0074
C7	no	1.889	0.0359	0.251	0.0216
* C8	YES	43.606	0.0147	87.896	0.9149
C9	no	2.176	0.0091	0.219	0.0033
**C10	YES	182.834	3.0062	340.852	0.1669
MIT-1	no	1.129	0.0095	1.443	0.0225
MIT-2	no	1.256	0.0163	1.615	0.0209
MIT-3	no	0.933	0.0204	0.537	0.0131
MIT-4	no	0.786	0.0067	1.027	0.0114

observe that distributions before and after present lower differences when the camera and the observed 2D tracks correspond to an area with especially peculiar geometric characteristics, but that the inclusion of object dynamics helps to correctly locate the camera in all cases. The proposed K-adventurers analysis has been performed on both estimated $q(\mathrm{E} \mid \Omega, \mathrm{M})$ and $p(\mathrm{E} \mid \Omega, \mathrm{M})$, and its conclusions (representative modes and their relative scores) have been included in the figure so as to indicate the ambiguity of each compared distribution.

As for the quantitative performance, Table 1 shows the absolute difference, in terms of both distance and orientation, between the real camera location and the estimation performed at the principal module, which corresponds to the best mode of the K-Adventurers analysis of the data. Due to the lack of comparable works in the literature, the table compares the obtained results with a baseline estimation method based on purely geometric point-wise consistency analysis and exhaustive search: a grid of 250×250 displacements and 720 camera orientations (i.e., resolution of 0.5 degrees) is checked, and each studied camera position is assigned a score consisting in the sum of the exponential of the signed Euclidean distances, normalized by W (Section 5.1.2), of all the corrected 2D points composing the track set Ω.

Table 1 presents results for the cameras of the synthetic and real databases, and not only indicates the accuracy reached by our proposal but also its ability to distinguish ambiguous settings and therefore detecting when the estimation should not be considered representative. In the tests, settings have been considered "unambiguous" whenever the weight associated to the best mode is above 0.9. The table shows that our proposal selected a wrong mode of the posterior distribution in two of the 14 tests: this is the case of cameras C4 and C10, corresponding to settings that fit several zones of the map. In other two tests, the correct mode was chosen but the provided estimate is clearly displaced from the actual location: this is the case of C1 and C8, corresponding to long straight areas of the map resulting in greatly dispersed modes. These two inherent ambiguous cases (four tests) are correctly detected by our approach as problematic.

The rest of cameras correspond to less ambiguous areas that have been correctly identified. In all these cameras, the distance between the estimation and the original camera is below 3.3 meters, and the orientation divergence is below 2 degrees. Our approach clearly outperforms the included baseline method in terms of reliability, providing satisfactory estimations in all cases. However, the baseline can provide slightly more accurate estimates in settings such as C7, C9 and MIT-3, where the observed tracks determine univocally the position of the camera without considering object dynamics.

7 Conclusions and Future Work

We proposed a framework for automatically estimating the location and orientation of a camera from observed tracks in its field of view. We use the information of the monitored scene in the form of a map that, to the best of our knowledge, has not been used before in this type of scenarios, and the adaptation of this information into a probabilistic dynamic model for objects to infer the location and orientation of the camera from observed trajectories. Experimental results show the capability of the proposal to analyze different camera settings, detecting those ambiguous situations where point estimation is not meaningful, providing satisfactory estimates in rest of cases.

Future work will aim at improving the accuracy of the resulting estimations for the unambiguous settings for example by using low-level details of object routes for optimization [41], [42]. Additionally, joint consistency between views could be exploited to disambiguate settings where individual cameras fit multiple map areas.

Appendices

Appendix A and Appendix B, available as online supplemental material, can be found on the Computer society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2013.243.

Acknowledgments

This work was partially supported by the Ministerio de Economía y Competitividad of the Spanish Government under project TEC2010-20412 (Enhanced 3DTV) and by the Artemis JU and UK Technology Strategy Board as part of the Cognitive & Perceptive Cameras (COPCAMS) project under GA number 332913. Also, Raúl Mohedano wishes to thank the Comunidad de Madrid for a personal research grant, which allowed him to do several parts of this work while visiting Queen Mary University of London.

References

[1] J. Deutscher, M. Isard, and J. MacCormick, "Automatic Camera Calibration from a Single Manhattan Image," *Proc. Seventh European Conf. Computer Vision (ECCV)*, pp. 175-205, 2002.

[2] P. Denis, J. Elder, and F. Estrada, "Efficient Edge-Based Methods for Estimating Manhattan Frames in Urban Imagery," *Proc. 10th European Conf. Computer Vision (ECCV)*, vol. 2, pp. 197-210, 2008.

[3] R.B. Fisher, "Self-Organization of Randomly Placed Sensors," *Proc. Seventh European Conf. Computer Vision (ECCV)*, pp. 146-160, 2002.

[4] I.N. Junejo, X. Cao, and H. Foroosh, "Autoconfiguration of a Dynamic Nonoverlapping Camera Network," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 37, no. 4, pp. 803-816, Aug. 2007.
[5] O. Javed, Z. Rasheed, O. Alatas, and M. Shah, "KNIGHT: A Real Time Surveillance System for Multiple and Non-Overlapping Cameras," *Proc. IEEE Int'l Conf. Multimedia and Expo (ICME)*, vol. 1, pp. 649-652, 2003.
[6] N. Krahnstoever and P.R.S. Mendonça, "Autocalibration from Tracks of Walking People," *Proc. British Machine Vision Conf.*, vol. 2, p. 107, 2006.
[7] B. Micusik, "Relative Pose Problem for Non-Overlapping Surveillance Cameras with Known Gravity Vector," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 3105-3112, 2011.
[8] M.B. Rudoy and C.E. Rohrs, "Enhanced Simultaneous Camera Calibration and Path Estimation," *Proc. 40th Asilomar Conf. Signals, Systems and Computers (ACSSC)*, pp. 513-520, 2006.
[9] Y. Caspi and M. Irani, "Aligning Non-Overlapping Sequences," *Int'l J. Computer Vision*, vol. 48, no. 1, pp. 39-51, 2002.
[10] M. Noguchi and T. Kato, "Geometric and Timing Calibration for Unsynchronized Cameras Using Trajectories of a Moving Marker," *Proc. Eighth IEEE Workshop Applications of Computer Vision*, p. 20, 2007.
[11] N. Anjum and A. Cavallaro, "Automated Localization of a Camera Network," *IEEE Intelligent Systems*, vol. 27, no. 5, pp. 10-18, Sept./Oct. 2012.
[12] A. Rahimi, B. Dunagan, and T. Darrell, "Simultaneous Calibration and Tracking with a Network of Non-Overlapping Sensors," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 187-194, 2004.
[13] V. John, G. Englebienne, and B. Krose, "Relative Camera Localisation in Non-Overlapping Camera Networks Using Multiple Trajectories," *Proc. 12th Int'l Conf. Computer Vision (ECCV)*, vol. 3, pp. 141-150, 2012.
[14] H. Andreasson, T. Duckett, and A.J. Lilienthal, "A Minimalistic Approach to Appearance-Based Visual SLAM," *IEEE Trans. Robotics*, vol. 24, no. 5, pp. 991-1001, Oct. 2008.
[15] D. Zou and P. Tan, "CoSLAM: Collaborative Visual SLAM in Dynamic Environments," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 354-366, Feb. 2013.
[16] C.F. Olson, "Probabilistic Self-Localization for Mobile Robots," *IEEE Trans. Robotics and Automation*, vol. 16, no. 1, pp. 55-66, Feb. 2000.
[17] S. Se, D. Lowe, and J. Little, "Global Localization Using Distinctive Visual Features," *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems*, vol. 1, pp. 226-231, 2002.
[18] A. Cesetti, E. Frontoni, A. Mancini, A. Ascani, P. Zingaretti, and S. Longhi, "A Visual Global Positioning System for Unmanned Aerial Vehicles Used in Photogrammetric Applications," *J. Intelligent and Robotic Systems*, vol. 61, no. 1-4, pp. 157-168, 2011.
[19] W.K. Leow, C.-C. Chiang, and Y.-P. Hung, "Localization and Mapping of Surveillance Cameras in City Map," *Proc. 16th ACM Int'l Conf. Multimedia*, pp. 369-378, 2008.
[20] O. Pink, "Visual Map Matching and Localization Using a Global Feature Map," *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1-7, 2008.
[21] G. Floros, B. van der Zander, and B. Leibe, "OpenStreetSLAM: Global Vehicle Localization Using OpenStreetMaps," *IEEE Int'l Conf. Robotics and Automation (ICRA)*, pp. 146-160, 2013.
[22] G. Vaca-Castano, A.R. Zami, and M. Shah, "City Scale Geo-Spatial Trajectory Estimation of a Moving Camera," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1186-1193, 2012.
[23] P. Lothe, S. Bourgeois, E. Royer, M. Dhome, and S. Naudet-Collette, "Real-Time Vehicle Global Localisation with a Single Camera in Dense Urban Areas: Exploitation of Coarse 3D City Models," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 863-870, 2010.
[24] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang, "Map-Matching for Low-Sampling-Rate GPS Trajectories," *Proc. ACM SIGSPATIAL Int'l Conf. Advances in Geographic Information Systems*, vol. 2, pp. 352-361, 2009.
[25] N.R. Velaga, M.A. Quddus, and A.L. Bristow, "Developing an Enhanced Weight-Based Topological Map-Matching Algorithm for Intelligent Transport Systems," *Transportation Research Part C: Emerging Technologies*, vol. 17, no. 6, pp. 113-140, 2009.
[26] I.N. Junejo, "Using Pedestrians Walking on Uneven Terrains for Camera Calibration," *Machine Vision and Applications*, vol. 22, no. 1, pp. 137-144, 2009.
[27] F. Lv, T. Zhao, and R. Nevatia, "Camera Calibration from Video of a Walking Human," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1513-1518, Sept. 2006.
[28] J.Y. Yen, "Finding the k Shortest Loopless Paths in a Network," *Management Science*, vol. 17, pp. 712-716, 1971.
[29] J. Hershberger, S. Suri, and A. Bhosle, "On the Difficulty of Some Shortest Path Problems," *ACM Trans. Algorithms*, vol. 3, no. 1, article 5, 2007.
[30] R. Mohedano and N. García, "Simultaneous 3D Object Tracking and Camera Parameter Estimation by Bayesian Methods and Transdimensional MCMC Sampling," *Proc. 18th IEEE Int'l Conf. Image Processing*, pp. 1873-1876, 2011.
[31] Z. Tu and S.-C. Zhu, "Image Segmentation by Data-Driven Markov Chain Monte Carlo," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 657-673, May 2002.
[32] Y. Rubner, C. Tomasi, and L.J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *Int'l J. Computer Vision*, vol. 40, no. 2, pp. 99-121, 2000.
[33] N. Anjum and A. Cavallaro, "Multifeature Object Trajectory Clustering for Video Analysis," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1555-1564, Nov. 2008.
[34] L. Devroye, *Non-Uniform Random Variate Generation*. first ed., Springer-Verlag, 1986.
[35] S.A. Golden and S.S. Bateman, "Sensor Measurements for Wi-Fi Location with Emphasis on Time-of-Arrival Ranging," *IEEE Trans. Mobile Computing*, vol. 6, no. 10, pp. 1185-1198, Oct. 2007.
[36] R. Mohedano and N. García, "Robust Multi-Camera 3D Tracking from Mono-Camera 2D Tracking Using Bayesian Association," *IEEE Trans. Consumer Electronics*, vol. 56, no. 1, pp. 1-8, Feb. 2010.
[37] Z. Khan, T. Balch, and F. Dellaert, "MCMC-Based Particle Filtering for Tracking a Variable Number of Interacting Targets," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1805-1819, Nov. 2005.
[38] X. Wang, X. Ma, and E. Grimson, "Unsupervised Activity Perception in Crowded and Complicated Scenes Using Hierarchical Bayesian Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 539-555, Mar. 2009.
[39] X. Wang, K. Tieu, and E. Grimson, "Correspondence-Free Activity Analysis and Scene Modeling in Multiple Camera Views," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 56-71, Jan. 2010.
[40] R.I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. second ed., Cambridge Univ. Press, 2004.
[41] D.I. Hastie and P.J. Green, "Model Choice Using Reversible Jump Markov Chain Monte Carlo," *Statistica Neerlandica*, vol. 16, no. 3, pp. 309-338, 2012.
[42] M. Hong, M.F. Bugallo, and P.M. Djurić, "Joint Model Selection and Parameter Estimation by Population Monte Carlo Simulation," *IEEE J. Selected Topics in Signal Processing*, vol. 4, no. 3, pp. 526-539, June 2010.

Raúl Mohedano received the ingeniero de telecomunicación degree (integrated BSc+MSc five years engineering program accredited by ABET) from the Universidad Politécnica de Madrid (UPM) and the licenciado en matemáticas degree (integrated BSc+MSc program) from the Universidad Complutense de Madrid (UCM), Madrid, Spain, in 2006 and 2012, respectively. Since 2007, he has been a member of the Grupo de Tratamiento de Imágenes (Image Processing Group) of the UPM, where he holds a personal research grant from the Comunidad de Madrid. His research interests are in the area of computer vision and probabilistic modeling.

Andrea Cavallaro received the PhD degree from the Swiss Federal Institute of Technology (EPFL), Lausanne, in 2002, and the laurea (summa cum laude) degree from the University of Trieste in 1996, both in electrical engineering. He is a professor of Multimedia Signal Processing and Director of the Centre for Intelligent Sensing at Queen Mary University of London, United Kingdom. He was a research fellow with British Telecommunications (BT) in 2004/2005 and received the Royal Academy of Engineering Teaching Prize in 2007; three Student Paper Awards on target tracking and perceptually sensitive coding at IEEE ICASSP in 2005, 2007, and 2009; and the Best Paper Award at IEEE AVSS 2009. He is the area editor for the *IEEE Signal Processing Magazine* and an associate editor for the *IEEE Transactions on Image Processing*. He is an elected member of the IEEE Signal Processing Society, Image, Video, and Multidimensional Signal Processing Technical Committee, and chair of its Awards committee. He served as an elected member of the IEEE Signal Processing Society, Multimedia Signal Processing Technical Committee, as an associate editor for the *IEEE Transactions on Multimedia* and the *IEEE Transactions on Signal Processing*, and as guest editor for seven international journals. He was General Chair for IEEE/ACM ICDSC 2009, BMVC 2009, M2SFA2 2008, SSPE 2007, and IEEE AVSS 2007. He was technical program chair of IEEE AVSS 2011, the European Signal Processing Conference (EUSIPCO 2008) and of WIAMIS 2010. He has published more than 130 journal and conference papers, one monograph on video tracking (2011,Wiley) and three edited books: *Multi-Camera Networks* (2009, Elsevier); *Analysis, Retrieval and Delivery of Multimedia Content* (2012, Springer); and *Intelligent Multimedia Surveillance* (2013, Springer).

Narciso García received the ingeniero de telecomunicación degree (five-year engineering program) in 1976 (Spanish National Graduation Award) and the doctor ingeniero de telecomunicación degree (PhD in communications) in 1983 (Doctoral Graduation Award), both from the Universidad Politécnica de Madrid (UPM), Spain. Since 1977, he has been a member of the faculty of the UPM, where he is currently a professor of signal theory and communications. He leads the Grupo de Tratamiento de Imágenes (Image Processing Group) of the Universidad Politécnica de Madrid. He has been actively involved in Spanish and European research projects, serving also as evaluator, reviewer, auditor, and observer of several research and development programmes of the European Union. He was a co-writer of the EBU proposal, base of the ITU standard for digital transmission of TV at 34-45 Mb/s (ITU-T J.81). He was the area coordinator of the Spanish Evaluation Agency (ANEP) from 1990 to 1992. He has been the general coordinator of the Spanish Commission for the Evaluation of the Research Activity (CNEAI) since 2011. He received the Junior and Senior Research Awards of the Universidad Politécnica de Madrid in 1987 and 1994, respectively. His professional and research interests are in the areas of digital image and video compression and of computer vision.