Neurocomputing 150 (2015) 155–162

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

An information geometric framework for the optimization on a discrete probability spaces: Application to human trajectory classification ☆

Jacinto C. Nascimento [a,*], Miguel Barão [b], Jorge S. Marques [a], João M. Lemos [c]

[a] ISR, Instituto Superior Técnico, Universidade de Lisboa, Portugal
[b] INESC-ID and Universidade de Évora, Portugal
[c] INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal

ARTICLE INFO

Article history:
Received 17 January 2014
Received in revised form
4 July 2014
Accepted 12 August 2014
Available online 8 October 2014

Keywords:
Fisher information metric
HMM
Human activity recognition
Natural gradient
Riemannian manifold
Surveillance

ABSTRACT

This paper presents an iterative algorithm using a information geometric framework to perform the optimization on a discrete probability spaces. In the proposed methodology, the probabilities are considered as points in a statistical manifold. This differs greatly regarding the traditional approaches in which the probabilities lie on a simplex mesh constraint. We present an application for estimating the switching probabilities in a space-variant HMM to perform human activity recognition from trajectories; a core contribution in this paper. More specifically, the HMM is equipped with a space-variant vector fields that are not constant but depending on the objects's localization. To achieve this, we apply the iterative optimization of switching probabilities based on the natural gradient vector, with respect to the Fisher information metric. Experiments on synthetic and realworld problems, focused on human activity recognition in long-range surveillance settings show that the proposed methodology compares favorably with the state-of-the-art.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

This paper presents a natural gradient method applied to the optimization on discrete (finite) probability spaces. Traditionally, this problem is tackled in a simplex probability constraint where standard gradient based methods are used. In this paper, we show that performing the optimization in a Riemannian space equipped with the Fisher metric provides several advantages over the standard methods. One of the advantages is that the Fisher metric is able to smoothly modify the gradient direction, so that it flows within the feasible region, i.e. the parameter space that satisfies all probability constraints. If some constraints become active, then the method behaves as the gradient projection method. Also, the Fisher metric exhibits a fast convergence since it behaves asymptotically as a Newton method. From the above, it will be shown that a formally correct interpretation of a natural gradient as the steepest-descent method is verified. With this approach the computational requirements are minimal: only marginally larger than the standard gradient; constant in time and space; and using rudimentary operations, i.e. additions and multiplications.

The novel contribution proposed in this paper is the application of the above framework in a new context: classification of human activities in far-field surveillance settings using the trajectories performed by pedestrians. Indeed, in the so-called far field scenarios, people are far from the camera, making it impossible to obtain detailed shape information and the system has to extract trajectories, or a rough shape description e.g., a bounding box or a coarse silhouette [7]. Models of typical trajectories may be estimated from training sets and then used to classify observed trajectories. This is one traditional problem that arises in outdoor surveillance systems and it will a focus of this paper. To model human trajectories in video sequences, we use a generative model of non-parametric vector fields proposed in [2]. The framework in [2] models the trajectories using a small set of vector or motion fields, estimated from observed trajectories. An advantage of using this model resides in its flexibility of modeling pedestrian's trajectories. More specifically, the trajectory is split into a sequence of segments, each of which generated by one vector field. Switching between models can occur at any point in the image but with

☆This work was supported by FCT Project PEst-OE/EEI/LA0009/2013, by projects "ARGUS"-PTDC/EEA-CRO/098550/2008, "PROBCONTROL"- PTDC/EEA-CRO/115038/2009 and by INESC-ID funding through PEst-OE/EEI/LA0021/2013.

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2014.08.074
0925-2312/© 2014 Elsevier B.V. All rights reserved.

probability that may depend on the spatial location. This provides a flexible tool to represent a wide variety of motion patterns. We present an expectation-maximization (EM) algorithm to learn the proposed model from sets of observed trajectories. The difference regarding the work in [2] is that here, the switching probabilities are estimated using the natural gradient instead of a projection simplex approach. These two methods, for computing the gradient, will be compared and the effectiveness of using the natural gradient will be illustrated.

2. Related work in human activity

Recognizing human activities in a quite diverse range of contexts and scenarios remains an up-to-date topic in image processing and computer vision communities. The goal is usually to interpret or classify human activities using tracked features. Related research in human activity analysis can be used in a wide variety of fields, such as intelligent environments [13], human machine interaction [14], surveillance [15,16], human computer interaction and sports analysis [17,18], to quote a few. Most of the work in this area falls into one of the two different settings, depending on which, the camera is close or far away to the person. In short range (SR) settings, the camera is close to the observed people, thus detailed information of human gestures, pose, gait can be computed. In long range or far field (FF) settings, the camera covers a wide area, thus no longer able to acquire a detailed type of information. Although, a large fraction of the related work on human activity recognition has been devoted to SR setup, we concentrate more to describe related work proposed in FF scenarios, that is the focus of the application presented herein, i.e. people are far from the camera and the trajectories are used as the information to perform classification/recognition.

In FF scenarios, it is usually impossible to obtain detailed descriptions of the observed persons, thus most methods rely only on the use of trajectories, taking for instance the center of the bounding box extracted by some region detection algorithm. Several trajectory analysis problems such as the one addressed herein, i.e. classification, have been addressed using pairwise similarity or dissimilarity measures between trajectories; these include Euclidean [4] and Hausdorff distances [5]. Because trajectories may have different lengths, techniques to face trajectories alignment have also been proposed. For instance, the use of dynamic time warping [6] or longest common subsequence [19] have been suggested to perform such comparisons. The class of approaches adopted in this paper models the trajectories as being produced by a probabilistic generative mechanism, usually an HMM or one of its variants [8–12]. These approaches have the key advantage of not requiring trajectory alignment or registration; moreover, they allow building a well grounded probabilistic inference formulation, based on which model parameters may be obtained from observed data. In that same class of approaches, in [20] the authors proposed a set of behavioral maps based on Markovian trajectory models, however, their application context is orthogonal to ours, since their goal is to improve tracking results by reconstructing full trajectories from fragments thereof.

All these contributions have been reported in [1], but in this paper, we provide a more comprehensive literature review, explanations and experimental results.

The paper is organized as follows. Section 3 describes the natural gradient proposed herein. Section 4 presents the generative model from which trajectory classification is performed. Section 5 describes how the generative model is learned with the EM algorithm using the natural gradient. Section 6 presents simulation results highlighting the superiority of the natural gradient and provides results using the proposed framework for human activity classification. Section 7 concludes the paper.

3. Discrete probability distributions in Riemannian space

This section provides detailed description of the proposed natural gradient. An usual premise is to assume that the probabilities lie on a simplex probability mesh. Contrasting with the above approach, we bring a new methodology, based on the information geometric framework [3,23,24], in which the probabilities are considered as points in a statistical manifold. We describe next how to parameterize the switching probabilities in the presented context.

One way to parameterize the probability mass function (p.m.f) of $p(x)$ defined over the set of p.m.f., $\mathcal{P}$, is to use the probabilities $\theta^k \stackrel{\text{def}}{=} Pr\{X=k\} = p(k)$, for $k=0,\ldots,K$. Since the probabilities satisfy the *partition of unity*, the probability θ^0 is defined as $\theta^0 \stackrel{\text{def}}{=} 1-\sum_{k=1}^{K}\theta^k$. This parameterization defined above provides a global coordinate system of $\mathcal{P}$, where θ^k are the coordinates.

3.1. The Fisher metric

The set $\mathcal{P}$ can be seen as a manifold, where each member $p \in \mathcal{P}$ is a p.m.f. and as an associated tangent space $T_p(\mathcal{P})$. One suited metric than can be introduced on the tangent space [3,23,24] uses the Fisher information matrix for defining the inner product on the manifold. The Fisher information matrix $\mathbf{G}_\theta$ has its entries defined as follows:

$$g_{ij}(\theta) \stackrel{\text{def}}{=} \mathbb{E}\left[\frac{\partial \log\, p(x)}{\partial\theta^i}\frac{\partial \log\, p(x)}{\partial\theta^j}\right] \tag{1}$$

It can be straightforwardly seen [3] that the components of the Fisher information matrix are given by

$$g_{ij}(\theta) = \frac{1}{1-\sum_{k=1}^{K}\theta^k} + \frac{\delta_{ij}}{\theta^i} \tag{2}$$

where δ_{ij} is the Kronecker delta function ($\delta_{ij}=1$ if $i=j$, $\delta_{ij}=0$ otherwise) and the corresponding Fisher information matrix $\mathbf{G}_\theta$ is given by

$$\mathbf{G}_\theta = \mathbf{I}_\theta + \left(1-\sum_{k=1}^{K}\theta^k\right)^{-1}\mathbf{1}\times\mathbf{1}^\top \tag{3}$$

where $\mathbf{I}_\theta$ is the $K\times K$ matrix having a diagonal structure with the i-th diagonal entry equal to $(\theta^i)^{-1}$ and $\mathbf{1}$ is a $K\times 1$ unit vector.

3.2. The natural gradient and its relation with Euclidean spaces

When performing optimization on $\mathcal{P}$, it is necessary to define a cost function F defined over $p\in\mathcal{P}$, i.e. $F(p)$. Since we are given the parameterization introduced in Section 3.1, it is possible to define a new function in these parameters, that is F_θ. Then, we can iterate over these parameters, until the convergence is reached.

Here, the parameter iteration that gives the gradient of F_θ in a arbitrary v-direction is the vector ∇F_θ such that the following equality holds:

$$\langle \nabla F_\theta, v\rangle = dF_\theta(v), \quad \forall v\neq 0. \tag{4}$$

Comparing with the standard Euclidean space, the inner product $\langle\cdot,\cdot\rangle$ (i.e. the gradient with respect to the Euclidean metric) is simply the product, and (4) becomes

$$\nabla F_\theta = \left[\frac{\partial F_\theta}{\partial\theta^1},\ldots,\frac{\partial F_\theta}{\partial\theta^K}\right]^\top \tag{5}$$

In the proposed context, i.e. in the Riemannian spaces, the inner product is defined by the metric tensor; the Riemannian (natural) gradient now becomes as an extension of (5) and given as follows:

$$\begin{aligned}\tilde{\nabla} F_\theta &= \mathbf{G}_\theta^{-1}\left[\frac{\partial F_\theta}{\partial \theta^1},\ldots,\frac{\partial F_\theta}{\partial \theta^K}\right]^\top \\ &= \mathbf{G}_\theta^{-1}\nabla F_\theta \end{aligned} \tag{6}$$

It can be easily seen that the final result (6) is the same as the standard gradient, if we replace the metric $\mathbf{G}$ by the identity matrix.

3.3. Optimization in the manifold

The optimization of the p.m.f. $p(x)\in\mathcal{P}$ can be formulated as a constrained optimization problem in $\mathbb{R}^K$, where the coordinates satisfy the *partition of unity* property. A well known optimization method is the steepest descent, which states that the parameters should follow negative gradient. The parameters update of this method can be written as

$$\theta^{(t+1)} = \theta^{(t)} - \eta^{(t)}\nabla F_\theta \tag{7}$$

where η is the step size. In Euclidean spaces, the steepest descent direction is given by ∇F_θ, whereas in Riemannian spaces the direction is given by natural gradient $\tilde{\nabla} F_\theta$ as defined in (6). Thus, for Riemannian the steepest descent method is re-defined as follows:

$$\theta^{(t+1)} = \theta^{(t)} - \eta^{(t)}\,\mathbf{G}_\theta^{-1}\nabla F_\theta. \tag{8}$$

3.4. Computational remarks of the natural gradient

One of the issues to take in to consideration when computing the natural gradient is that of inverting the matrix $\mathbf{G}_\theta$, since it can lead to a numerical problems. The metric given in (3) for a given discrete p.m.f. is badly conditioned and numerical problems may arise. Thus, the goal here is precisely to avoid the inversion in (6). Following [3], we start by rewriting the Fisher information matrix as

$$\mathbf{G}_\theta = \mathbf{I}_\theta + \mathbf{1}\left(1-\sum_{k=1}^{K}\theta^k\right)^{-1}\mathbf{1}^\top \tag{9}$$

where $\mathbf{I}_\theta$ is the $K\times K$ entity matrix with the i-th diagonal entry equal to $(\theta^i)^{-1}$ and $\mathbf{1}$ is a $K\times 1$ unit vector (as in (3)). Using the Woodbury identity, the matrix inverse $\mathbf{G}^{-1}$ can be expressed as

$$\mathbf{G}_\theta^{-1} = \mathbf{I}_\theta^{-1} - \mathbf{I}_\theta^{-1}\,\mathbf{1}\left(\mathbf{1}^\top\mathbf{I}_\theta^{-1}\mathbf{1} + \left(1-\sum_{k=1}^{K}\theta^k\right)\right)^{-1}\mathbf{1}^\top\mathbf{I}_\theta^{-1}, \tag{10}$$

which can be further simplified to

$$\mathbf{G}_\theta^{-1} = \mathbf{I}'_\theta - \boldsymbol{\theta}\boldsymbol{\theta}^\top \tag{11}$$

where, $\mathbf{I}'_\theta$ is the $K\times K$ matrix having an identity structure with the i-th diagonal entry equal to (θ^i) and $\boldsymbol{\theta} = [\theta^1,\ldots,\theta^K]^\top$.

From (11) it can be easily seen that the information matrix can be computed avoiding its inversion when computing the natural gradient.

Taking (11) and (6), we can write the natural gradient, $\tilde{\nabla} F_\theta$ in terms of the standard gradient ∇F_θ, i.e.

$$\tilde{\nabla} F_\theta = \boldsymbol{\theta}\circ\nabla F_\theta - \boldsymbol{\theta}(\boldsymbol{\theta}\cdot\nabla F_\theta) \tag{12}$$

where the notation $\circ$ stands for the Hadamard product. From (12) we conclude that it s not required to explicitly compute the entire matrix $\mathbf{G}_\theta^{-1}$ to compute the product $\mathbf{G}_\theta^{-1}\nabla F_\theta$ as in (6).

Finally, one has to decide when the optimization iterative procedure reaches its final solution. A natural approach would be to compare the components of the gradient vector against a pre-defined threshold τ. This is however a misleading choice, for the following two reasons: first, this approach does not take into consideration the geometric nature of the underlying space, second, the components of the gradient vector have different behaviors depending on the curvature of F_θ. To tackle the above issues, the norm of the natural gradient is used and can be developed as follows:

$$\begin{aligned}\|\tilde{\nabla} F_\theta\| &= \tilde{\nabla} F_\theta^\top\,\mathbf{G}_\theta\tilde{\nabla} F_\theta \\ &= \nabla F_\theta^\top\,\mathbf{G}_\theta^{-1}\nabla F_\theta \\ &= \nabla F_\theta^\top\,\tilde{\nabla} F_\theta \end{aligned} \tag{13}$$

where the second and third equalities comes from the result in (6). Thus, the iterative procedure stops when the rate change of F_θ is below a pre-defined threshold τ.

4. Multiple vector fields for describing trajectories

This section briefly revises the framework proposed in [2] for modeling trajectories that take place in far-field surveillance scenarios. We aim to demonstrate that the natural gradient is an efficient methodology to estimate the parameters (i.e. transition probabilities) of the generative model described in the next section.

4.1. Generative motion model

We will denote the set of vector motion fields as $\mathcal{T} = \{\mathbf{T}_1,\ldots,\mathbf{T}_K\}$, with $\mathbf{T}_{k_t}:\mathbb{R}^2\to\mathbb{R}^2$, for $k_t\in\{1,\ldots,K\}$. The generative motion model of the trajectory is given as

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{T}_{k_t}(\mathbf{x}_{t-1}) + \mathbf{w}_t, \quad t=2,\ldots,L, \tag{14}$$

where $\mathbf{w}_t\sim\mathcal{N}(0,\sigma^2_{k_t}\mathbf{I})$ is white Gaussian noise with zero mean and variance $\sigma^2_{k_t}$ (which may be different for each field), and L is the number of points in the trajectory. Also, we will assume that the sequence of active fields $\mathbf{k} = \{k_1,\ldots,k_L\}$ is modeled as a realization of a first order Markov process with space varying transition probabilities. This model allows the switching to depend on the object localization, thus having $P(k_t = j|k_{t-1}=i,\mathbf{x}_{t-1}) = \mathbf{B}_{ij}(\mathbf{x}_{t-1})$, where $\mathbf{B}:\mathbb{R}^2\to\mathbb{R}^{K\times K}$ is a field of stochastic matrices. The matrix $\mathbf{B}$ can also be seen as a set of K^2-dimensional fields with values in $[0,1]$ s.t. $\sum_j B_{ij}(\mathbf{x}_t) = 1$, for any $\mathbf{x}_t$ and any i.

The joint distribution of a trajectory $\mathbf{x}$ and its underlying sequence of active fields $\mathbf{k}$, under the model parameters $\boldsymbol{\Theta} = (\mathcal{T},\mathbf{B},\boldsymbol{\Sigma})$, is given by

$$p(\mathbf{x},\mathbf{k}|\boldsymbol{\Theta}) = p(\mathbf{x}_1)P(k_1)\prod_{t=2}^{L}p(\mathbf{x}_t|\mathbf{x}_{t-1},k_t)p(k_t|k_{t-1},\mathbf{x}_{t-1}). \tag{15}$$

From (15), we see that $p(k_t|k_{t-1},\mathbf{x}_{t-1})$ is a function of $\mathbf{B}$, $p(\mathbf{x}_t|\mathbf{x}_{t-1},k_t)$ is a function of $\mathcal{T}$ and $\boldsymbol{\Sigma}$, and $p(\mathbf{x}_t,k_t|\mathbf{x}_{t-1},k_{t-1})$ is a function of $\mathcal{T}$, $\mathbf{B}$, and $\boldsymbol{\Sigma}$.

As in [2] both of the fields and transition matrices $(\mathcal{T},\mathbf{B})$ are modeled in a non-parametric way. More specifically, they are defined at the nodes of a regular grid. To obtain the velocity fields and switching probability fields, we interpolate the vectors $\mathbf{t}_k^{(n)}$ and matrices $\mathbf{b}^{(n)}$ defined at the nodes of the grid as follows:

$$\mathbf{T}_k(\mathbf{x}) = \sum_{n=1}^{N}\mathbf{t}_k^{(n)}\phi_n(\mathbf{x}), \quad \mathbf{B}(\mathbf{x}) = \sum_{n=1}^{N}\mathbf{b}^{(n)}\phi_n(\mathbf{x}) \tag{16}$$

where $\phi_n(\mathbf{x}):\mathbb{R}^2\to\mathbb{R}$, for $n=1,\ldots,N$ is a set of scalar basis functions. Given the image domain $\mathcal{D} = [0,1]^2$, a discretization is performed using an uniform grid with step Δ. The contribution of node n,

centered at $\mathbf{u}_n=(u_n^1,u_n^2)$, to the interpolation (16) is given by

$$\phi_n(\mathbf{x})=\begin{cases}|x^1-u_n^1|\cdot|x^2-u_n^2|/\Delta^2 & \text{if } |x^1-u_n^1|<\Delta \text{ and } |x^2-u_n^2|<\Delta,\\ 0 & \text{otherwise.}\end{cases}$$

5. Learning the model with EM algorithm

Here, we detail how the model parameters $\boldsymbol{\Theta}=(\mathcal{T},\mathbf{B},\boldsymbol{\Sigma})$ are learned. More specifically, how the motion fields $\mathcal{T}=\{\mathbf{T}_1,\ldots,\mathbf{T}_K\}$, the field of the stochastic matrices $\mathbf{B}$ and the noise variances $\boldsymbol{\Sigma}=\{\sigma_1^2,\ldots,\sigma_K^2\}$ are learned from a set of S independent observed trajectories $\mathcal{X}=\{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(S)}\}$, where $\mathbf{x}^{(s)}=(\mathbf{x}_1^{(s)},\ldots,\mathbf{x}_{L_s}^{(s)})$ is the s-th observed trajectory. Since we assume that the sequence of active models $\mathcal{K}=\{\mathbf{k}^{(1)},\ldots,\mathbf{k}^{(S)}\}$ are missing, we apply the EM algorithm to find a *marginal maximum a posteriori* (MMAP) estimate of $\boldsymbol{\Theta}$; formally the estimate is given by

$$\begin{aligned}\widehat{\boldsymbol{\Theta}}&=\arg\max_{\boldsymbol{\Theta}}\sum_{\mathcal{K}}p(\mathcal{X},\mathcal{K}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})\\&=\arg\max_{\boldsymbol{\Theta}}\sum_{\mathcal{K}}\prod_{s=1}^{S}p(\mathbf{x}^{(s)},\mathbf{k}^{(s)}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})\end{aligned}\tag{17}$$

where each factor $p(\mathbf{x}^{(s)},\mathbf{k}^{(s)}|\boldsymbol{\Theta})$ has the form given in (15), the sum over $\mathcal{K}$ has $K^{(\sum_s L_s)}$ terms and $p(\boldsymbol{\Theta})=p(\mathcal{T})p(\mathbf{B})p(\boldsymbol{\Sigma})$ is some prior.

5.1. The complete log-likelihood

The EM algorithm aims at computing (the E-step) the expectation of the complete log-likelihood which is given by

$$\begin{aligned}Q(\boldsymbol{\Theta};\widehat{\boldsymbol{\Theta}})&\equiv\mathbb{E}\left[\log\ p(\mathcal{X},\mathcal{K}|\boldsymbol{\Theta})|\mathcal{X},\widehat{\boldsymbol{\Theta}}\right]\\&=\underbrace{\sum_{s=1}^{S}\sum_{t=2}^{L_s}\sum_{l=1}^{K}\overline{w}_{t,l}^{(s)}\log\ \mathcal{N}(\mathbf{x}_t^{(s)}|\mathbf{x}_{t-1}^{(s)}+\mathbf{T}_l(\mathbf{x}_{t-1}^{(s)}),\sigma_l^2\mathbf{I})}_{\overline{\mathcal{A}}(\mathcal{X},\mathcal{K})}\\&+\underbrace{\sum_{s=1}^{S}\sum_{t=2}^{L_s}\sum_{l=1}^{K}\sum_{g=1}^{K}\overline{w}_{t,g,l}^{(s)}\log\ B_{g,l}(\mathbf{x}_{t-1}^{(s)})}_{\overline{\mathcal{B}}(\mathcal{X},\mathcal{K})},\end{aligned}\tag{18}$$

where $\overline{w}_{t,l}^{(s)}=P[w_{t,l}^{(s)}=1|\mathbf{x}^{(s)},\widehat{\boldsymbol{\Theta}}]$, and $\overline{w}_{t,g,l}^{(s)}=P[w_{t,g,l}^{(s)}=1|\mathbf{x}^{(s)},\widehat{\boldsymbol{\Theta}}]$ which are obtained by a modified forward–backward procedure [21].

The M-step maximizes the Q-function in (18) with respect to the model parameters $\boldsymbol{\Theta}$. The maximization with respect to the motion vector fields $\mathcal{T}$ and noise variances $\boldsymbol{\Sigma}$ (the term $\overline{\mathcal{A}}(\mathcal{X},\mathcal{K})$ in (18)) is straightforward and it follows the same strategy as in [2]. The novelty resides how we optimize the term $\overline{\mathcal{B}}(\mathcal{X},\mathcal{K})$ in (18)), i.e. the transition probabilities. To accomplish this, we first take the transition matrix in (16). Notice that each component $\mathbf{b}^{(n)}\in\mathbb{R}^{K\times K}$ is a stochastic matrix at any location $\mathbf{x}$, satisfying the constraint

$$\sum_{k=1}^{K}b_{p,k}^{(n)}=1.\tag{19}$$

Given the formulation in (16), the problem of estimating $\mathbf{B}$ is the same as estimating $(\mathbf{b}^{(1)},\ldots,\mathbf{b}^{(N)})$, by maximizing the objective function $Q(\boldsymbol{\Theta};\widehat{\boldsymbol{\Theta}})$ under the constraint (19). Inserting (16) into (18) the objective function becomes

$$Q(\boldsymbol{\Theta};\widehat{\boldsymbol{\Theta}})=\overline{\mathcal{A}}(\mathcal{X},\mathcal{K})+\sum_{s=1}^{S}\sum_{t=2}^{L_s}\sum_{l=1}^{K}\sum_{g=1}^{K}\sum_{n=1}^{N}\overline{w}_{t,g,l}^{(s)}\log\ b_{g,l}^{(n)}\phi_n(\mathbf{x}),\tag{20}$$

Now, we derive (20) with respect to $b_{p,k}^{(n)}$, where $b_{p,k}^{(n)}$, stands for the p-th line, k-column at n-node. Denoting the derivative of all transition matrices in the grid as $\nabla\mathbf{B}$, this derivative is the term ∇F_θ in (12) and $\mathbf{B}$ is the term θ in (12). Thus, the natural gradient provides the following update rule:

$$\mathbf{B}^{(t+1)}=\mathbf{B}^{(t)}+\eta\left(\mathbf{B}^{(t)}\circ\nabla\mathbf{B}^{(t)}-\mathbf{B}^{(t)}(\mathbf{B}^{(t)}\cdot\nabla\mathbf{B}^{(t)})\right)\tag{21}$$

where η is the step size.

6. Experimental results

This section reports experimental results using the proposed approach. In the first part of the experiments, we report simulation results of the framework described in Section 3. In the second part of the experiments we focus on the learning the generative model in (14) to perform activity classification from the trajectories performed by pedestrians. In the first part of the results (simulation), we perform a qualitative comparison between the natural gradient and the standard gradient. In the second part of the experiments (trajectory classification) we perform a comparison between the natural gradient and the projection simplex algorithm [22], where a fast gradient projections approach with ℓ_1 domain constraints is proposed.

6.1. Simulation results

To illustrate the application of the natural gradient, we start by presenting two examples where the goal is to show the superiority of the natural gradient when compared to the standard gradient.

6.1.1. Example 1

In the first example, we are given a p.m.f. $p(x)$ to minimize the $K-L$ divergence, $\mathcal{D}(p||q)$ where $q(x)$ is a target p.m.f. Although, the

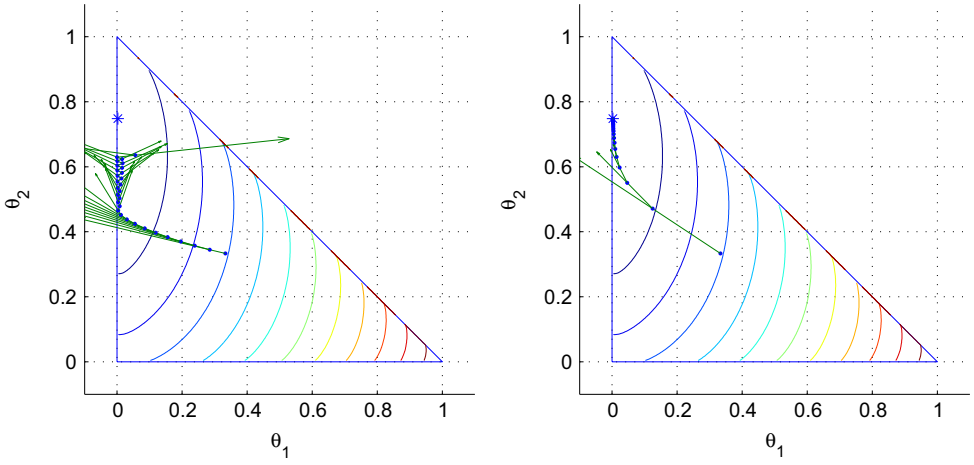

Fig. 1. Optimization of the $K-L$ divergence $\mathcal{D}(p||q)$ using the standard (left) and natural (right) gradient.

solution is known, i.e. $p(x)=q(x)$, this problem is useful since it corresponds to a scenario in which both gradient and natural gradient can be easily be compared.

Denoting $\mathbf{p}=[\theta^0\ \theta^1\ \ldots\ \theta^K]^\top$, with the constraint $\theta^0 \stackrel{\text{def}}{=} 1-\sum_{k=1}^{K}\theta^k$, the standard gradient of the $K-L$ divergence can be written as

$$\nabla\mathcal{D}=[-\mathbf{1}\ \mathbf{I}]\,(\log\,\mathbf{p}-\log\,\mathbf{q}) \tag{22}$$

where $\mathbf{1}$ is a $K\times 1$ vector and $\mathbf{I}$ is a $K\times K$ identity matrix. We now assume that the target p.m.f. is given as $\mathbf{q}=[0.2494\ 0.0025\ 0.7481]^\top$, and $\mathbf{p}=[\frac{1}{3}\ \frac{1}{3}\ \frac{1}{3}]^\top$ as an initial guess. Fig. 1 illustrates the behavior of the standard (left) and natural (right) gradient using a constant step size $\eta=0.01$. From this figure (left), it is possible to observe the following two limitations of the standard gradient. When the curvature exhibits significant and different directions a small step η is required. In this situation an undesirable slow convergence is obtained. On the contrary, if the step size η is large, then instability occurs. Fig. 1 (right) shows the superiority of the natural gradient, where the two above limitations are now overpassed. Here the step-size is set to $\eta=0.18$ and it is seen how fast is the convergence in the proposed technique. The reason for this behavior is that the optimization using a Fisher metric behaves asymptotically as a Newton method and exhibits fast convergence rate near the optimum, i.e. $p=q$.

6.1.2. Example 2

In the second example, the $K-L$ divergence $\mathcal{D}(p||q)$ is again minimized, as before, however we assume that p is generated by the joint p.m.f. $p(x,y)=p(y|x)p(x)$, where $p(y|x)$ is the likelihood assumed to be known and $p(x)$ is a free prior, i.e. we have $p(x,y)=p(y|x)p_\theta(x)$. The optimization is performed in the prior $p_\theta(x)$.

The gradient of the cost function $\mathcal{D}(p(x,y)||q(x,y))$ is as follows:

$$\frac{\partial\mathcal{D}}{\partial\theta^i}=\sum_{x,y}p(y|x)\log\frac{p(y|x)p(x)}{q(x,y)}\frac{\partial p(x)}{\partial\theta^i} \tag{23}$$

Adopting a matrix notation, we next describe how this example can be implemented. Thus, $\mathbf{p}_X$ denotes a $K\times 1$ vector of probabilities $p(x)$; $\mathbf{P}_{Y|X}$ is a $L\times K$ matrix; $\mathbf{P}_{Y,X}$ and $\mathbf{Q}_{Y,X}$ are the joint probability distributions of $p(x,y)$ and $q(x,y)$ with $L\times K$ size, respectively. Using the above formulation and similarly to (22) the standard gradient is given by

$$\nabla\mathcal{D}=[-\mathbf{1}\ \mathbf{I}]\,\mathbf{T}_{X,Y}\,\mathbf{1} \tag{24}$$

where $\mathbf{T}_{X,Y}$ is given by

$$\mathbf{T}_{Y,X}=\mathbf{p}_{Y|X}\circ(\log\,\mathbf{P}_{Y,X}-\log\,\mathbf{Q}_{Y,X}) \tag{25}$$

and $\mathbf{P}_{Y,X}$ given by

$$\mathbf{P}_{Y,X}=\mathbf{P}_{Y|X}\circ\mathbf{1}\,\mathbf{p}_X^\top \tag{26}$$

with $\mathbf{1}$ a $L\times 1$ vector and $\circ$ a elementwise product.

Recall that to compute the natural gradient, we have to compute the matrix $\mathbf{G}^1$ that is given in (12).

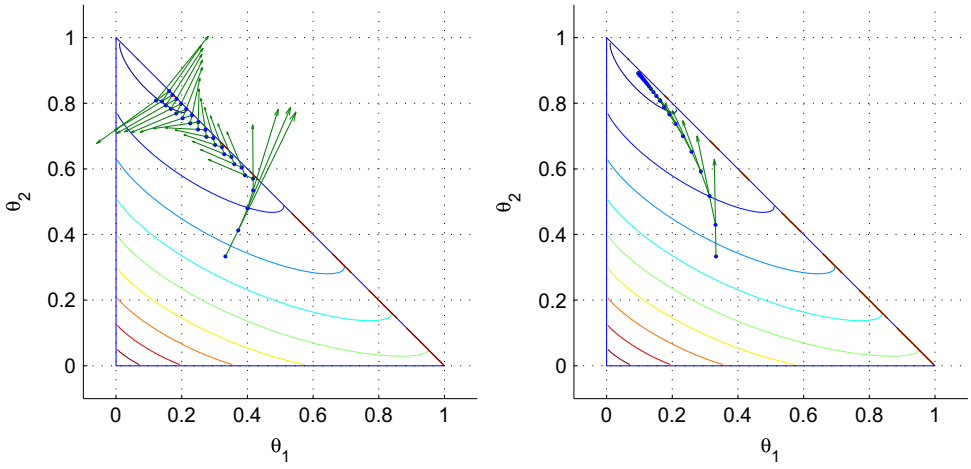

Fig. 2. Optimization of the $K-L$ divergence $\mathcal{D}(p||q)$ using the standard (left) and natural (right) gradient where p is assumed to be jointly distributed.

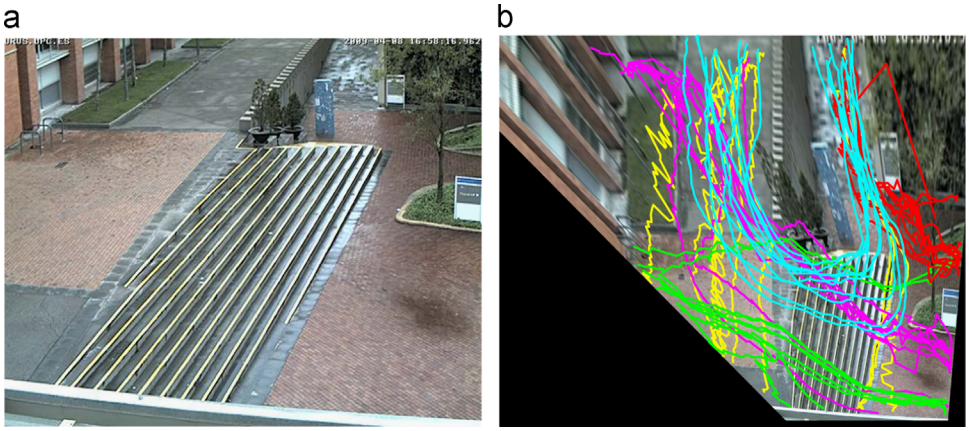

Fig. 3. (a) Scenario of the campus in Barcelona, (b) homography of the image in (a) superimposed with the homography of the trajectories. Each color refers to a different activity (see text). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

In this example the $\mathbf{P}_{Y,X}$, $\mathbf{P}_{Y,X}$ have dimensions set to 3×2 (i.e. $K=2$, $L=3$). Fig. 2 shows the evolution of the parameters for standard (left) and the proposed natural (right) gradient. It can be seen that the cost $\mathcal{D}(p||q)$ does not converge to zero since there is no distribution $p(x)$ such that $p(x)p(y|x)=q(x,y)$, as was the case in the first example.

6.2. Trajectory classification

This section presents an example that is related to the generative model of vector fields presented in Section 4. More specifically, we use the proposed natural gradient for estimating the switching probabilities in the generative model, as detailed in Section 5.1.

In the trajectory classification context, we assume that we have a predefined number of activities and that we have a subset of trajectories from each of these activity classes, $\mathcal{X}^1,\ldots,\mathcal{X}^A$, where A is the admissible number of activities. For each activity class a, we denote the corresponding fields as $\theta^a=(\mathcal{T}^a,\mathbf{B}^a,\mathbf{\Sigma}^a)$, with $a=1,\ldots,A$. In the example presented, several activity classes share some of the vector fields. However, the switching probabilities, i.e. the transitions among the motion models within a trajectory, are specific and different for each activity. In this section we will illustrate the usefulness of the proposed approach to tackle the above task. Also, we provide a comparison with an efficient method for projection onto the probabilistic simplex proposed in [22].

6.2.1. Real data

This section reports the performance of the proposed approach to pedestrian activity classification in the context of far-field surveillance scenarios. The images were obtained from a network camera located at the campus of the Universitat Politécnica de Catalunya (UPC) Barcelona. After several hours of recording, we observed that the pedestrians performed common paths in the scenario. Thus, it was possible to collect the most common trajectories and organize them into several classes.

Before estimating the parameters of the generative model for the activities, i.e. $\mathbf{\Theta}_K^{(a)}=(\mathcal{T},\mathbf{B}^{(a)},\mathbf{\Sigma})$, we pre-processed all the trajectories computing the homography. The reason behind is that the vector fields obtained in this way are more distinguishable being more easily to estimate. Fig. 3(a) shows an image of the scenario of the campus considered in the experiments. Fig. 3(b) shows the corresponding homography of the previous image, as well as the homography of the trajectories considered in the scenario.

Before applying the proposed model to estimate a set of motion fields, we need to extract the trajectories from the video sequences by tracking the pedestrians. For that purpose, we used the Lehigh omnidirectional tracking system (LOTS) [25] to detect regions, followed by region association. Region association works as follows: a pair of regions (A_t,B_{t+1}) detected in consecutive frames is associated if B_{t+1} is the only region in the second frame that overlaps (above a given threshold) with A_t and vice versa. This can be interpreted as mutual favorite pairing. If the obtained associated trajectories have some small gaps, we manually edit to correct wrong or missing connections. The trajectories are then projected onto a view orthogonal to the ground plane (the so called bird's eye view) to enforce viewpoint invariance. This is done using a projective transformation (homography) from the image onto a plane parallel to the ground. The number of the trajectories are 270 for the scenario considered.

The trajectory classes are organized as follows: *walking and stepping up the stairs* (red), *walking along* (green), *crossing and stepping up the stairs* (yellow), *pass diagonally up* (magenta) and *turning the Campus* (cyan). Recall that, as above mentioned, this is a difficult example since the motions (i.e. vector fields) are similar among classes. Only the transitions may contain specific information regarding each class-type of trajectories. Precisely the goal of the natural gradient, that is accurately estimate the switching probabilities, a crucial information contained in the trajectories to achieve a good performance at classifying human activities from trajectories.

The experimental evaluation conducted is different from [1]. In the previous work we considered a separate validation or selection set to determine the best model order. Also, we did not considered the variability of the solution regarding different initializations of the EM. Thus, the following issues are taken into consideration in this paper:

- *Number of motion fields*: To determine the appropriate number of vector fields, we varied the number of motion models in the interval $K\in\{1,\ldots,6\}$. The best model order is chosen such that maximizes the performance classification accuracy.
- *Initializations of the EM*: Given the variability of the EM estimates with the initializations, we perform eight different initialization of the EM (i.e. eight runs of the EM), to obtain the statistics of the results.
- *Cross validation*: Since the number of trajectories cannot be indefinitely generated, we perform a 5-fold cross validation to obtain the performance evaluation. These splitting between the training and test sets is done randomly, but guaranteeing roughly balanced sets in terms of classes.
- *Range of the step size* η: we considered the range of this parameter $\eta\in\{1\times 10^{-3},\ldots,1\times 10^{-9}\}$ and present the final statistics for the best value η for which both approaches exhibit higher accuracy in the trajectory classification.

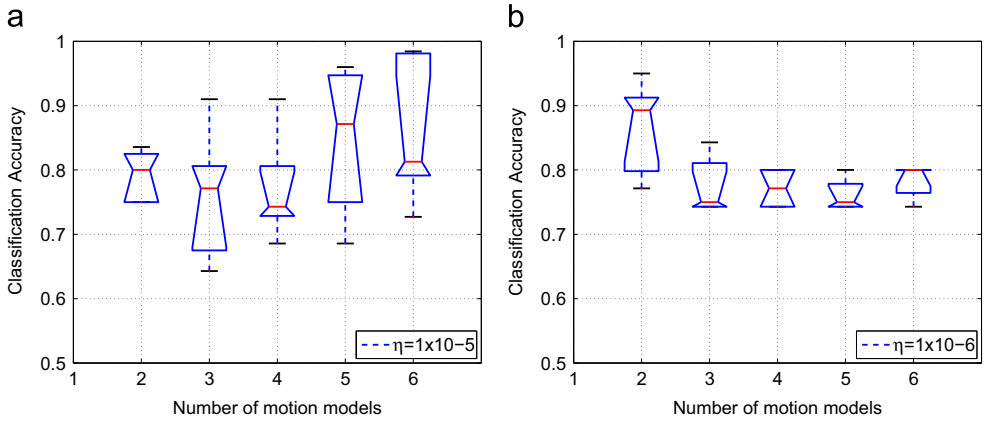

Fig. 4. Comparison of the two methodologies (a) using the projection simplex and (b) varying the step size η and the number of motion fields K.

Summarizing the procedure: for each number of motion fields $K \in \{1,\ldots,6\}$ we perform the classification, for eight runs of the EM in each fold $\mathcal{F} \in \{1,\ldots,5\}$ in which we take the best run. We repeat this procedure for each value of the step size η to obtain the statistics. Recall that with the above procedure, we are assuming that all the activities *share* the same vector fields, i.e. the class specific models are $\boldsymbol{\Theta}_K^{(a)} = (\mathcal{T}, \mathbf{B}^{(a)}, \boldsymbol{\Sigma})$, for $a \in \{1,\ldots,A\}$, where only the switching matrices differ among the classes, and K is the number of (shared) vector fields.

Fig. 4 discriminates the accuracy in terms of trajectory classification among the considered classes varying the number of motion models for the ranges of K. The best values of the step-size are shown for both of the methodologies. We illustrate the results for the best initialization of the EM in the folds. Results concerning the use of $K=1$ are not shown, since both methodologies do not provide acceptable performance. This happens since one single model does not suffice to represent the variability of the motions present in the activities of the scenario.

Fig. 4 clearly shows that both approaches (i.e. using the natural gradient and the projection simplex) are remarkably competitive providing high accuracy rates.

We also compute the running time figures of both methods for computing the switching matrix. In the test conditions, we used the best step size η for each method as shown in Fig. 4.[1] We perform two experiments varying the number of motion fields. More specifically, the experiments comprise $K=2$ and $K=6$, corresponding to the less and most expensive computational scenarios, respectively. Thus, for the first situation ($K=2$) we obtained $t \approx 0.13$ s. for the projection simplex and $t \approx 0.04$ s. for the natural gradient. For the second experiment ($K=6$) we obtained $t \approx 0.35$ s. for the projection simplex and $t \approx 0.08$ s. for the proposed approach. One of the reasons for the fastest computation of the proposed approach is that we no longer perform the projection to estimate the transition matrix. These running time figures are obtained in the M-step and only for the computation of $\mathbf{B}$.[2] These results support our claims in the Introduction, where we state that the Fisher metric exhibits faster convergence behaving as a Newton method.

From the results obtained in Fig. 4 we conclude that the natural gradient exhibits better stability in the results when varying the number of motion fields in terms of both classification accuracy and covariance assigned to these scores. Notice however that the results provided in Fig. 4 are obtained when sharing the motion fields among different type of trajectory classes, not exploring all the capabilities of the framework proposed herein. This means that the number of the motion fields is the same for all the activity classes. Another possibility is to assume that each trajectory-class has its own complexity, meaning that each activity may require a different number of vector fields, thus having a different complexity. Under this scenario, the model for each activity/class is independently estimated from the subset of the training data from that class, i.e. we independently learn class-specific models. Although, this approximation is more complex, since it requires different classification for each motion model configurations, still it is possible to be explored. Also note that in the proposed approach the step-size is fixed. A different strategy is to have variable step-size which can also enhance the capabilities of the framework.

[1] Notice however that the time we report here is independent of η.

[2] Recall that in this experiment we do not provide the time expended by the overall M-step, since we estimate the triplet $(\mathcal{T}, \mathbf{B}^{(a)}, \boldsymbol{\Sigma})$ that can computed independently. Thus, in the experiment reported it is possible to take only the time needed for computing $\mathbf{B}$.

7. Conclusions

We presented a natural gradient approach applied to the optimization of the Kullback Leibler ($K-L$) divergence. The above methodology contrasts with traditional approaches, in the sense that the probabilities do not lie on a simplex probability mesh, but considered as points in the manifold instead. Simulation results have shown that the proposed method converges faster and attained better accuracy comparing with the standard gradient method. Also, it has been shown that this framework is suited to estimate space varying switching probabilities in the generative model in the context of human activity recognition in far field surveillance settings. The results shown allow us to conclude that the proposed algorithm compares favorably with sate of the art methods applied in this new context of application. Further work will include some of the directions mentioned in the previous section that allow to enlarge the framework both in terms of practical and theoretical contexts.

References

[1] J.C. Nascimento, M. Barão, J.M. Lemos, J.S. Marques, Efficient optimization algorithm for space-variant switching of vector fields, in: Iberian Conference on Pattern Recognition and Image Analysis (IbPria), 2013, pp. 79–88.

[2] J. C. Nascimento, M. A. T. Figueiredo and J. S. Marques, Activity recognition using mixture of vector fields, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 22, NO. 5, MAY 2013. http://dx.doi.org/10.1109/TIP.2012.2226899.

[3] M. Barão, J.M. Lemos, An efficient Kullback-Leibler optimization algorithm for probabilistic control design, in: Mediterranean Conference on Control and Automation, 2008, pp. 198–203.

[4] Z. Fu, W. Hu, T. Tan, Similarity based vehicle trajectory clustering and anomaly detection, in: Proceedings of the IEEE International Conference on Image Processing, 2005, pp. 602–606.

[5] X. Wang, K. Tieu, and E. Grimson, Learning semantic scene models by trajectory analysis, in: Proceedings of the European Conference on Computer Vision, 2006, pp. 110–123.

[6] M. Pierobon, M. Marcon, A. Sarti, S. Tubaro, Clustering of human actions using invariant body shape descriptor and dynamic time warping, in: Proceedings of the IEEE Conference on Advance Video Signal Based Survey, 2005, pp. 22–27.

[7] N. Robertson, I. Reid, A general method for human activity recognition in video, Comput. Vis. Image Understand. 104 (2) (2006) 232–248.

[8] Y. Du, F. Chena, W. Xua, W. Zhanga, Activity recognition through multi-scale motion detail analysis, Neurocomputing 71 (18) (2008) 3561–3574.

[9] T. Duong, H. Bui, D. Phung, S. Venkatesh, Activity recognition and abnormality detection with the switching hidden semi-Markov model, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 838–845.

[10] J.C. Nascimento, M.A.T. Figueiredo, J.S. Marques, Independent increment processes for human motion recognition, Int. J. Comput. Vis. Image Understand. 109 (2) (2008) 126–138.

[11] N. Oliver, B. Rosario, A. Pentland, A Bayesian computer vision system for modeling human interactions, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 831–843.

[12] Q. Shi, L. Wang, Disciminative human action segmentation and recognition using semi-Markov model, in: Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, 2008, pp. 1–6.

[13] D. Hu, S. Pan, V. Zheng, N. Liu, Q. Yang, Real world activity recognition with multiple goals, in: ACM 10th International Conference Series, 2008, pp. 30–39.

[14] M. Pantic, A. Pentland, A. Nijholt, T. Huang, Human computing and machine understanding of human behavior: a survey, in: 8th International Conference on Artificial Intelligence on Human Computer, 2007, pp. 239–248.

[15] W. Hu, T. Tan, L. Wang, S. Maybank, A survey on visual surveillance of object motion and behaviors, IEEE Trans. Syst. Cybern. C, Appl. Rev. 34 (3) (2004) 334–352.

[16] L. Weilun, H. Jungong, P. With, Flexible human behavior analysis framework for video surveillance, Int. J. Digit. Multimed. Appl. (9) (2010) 1687–7578.

[17] P. Barr, J. Noble, R. Biddle, Video game values: human-computer interaction and games, Interact. Comput. 19 (2) (2007) 180–195.

[18] M. Perse, M. Kristan, S. Kovacic, G. Vuckovic, J. Persa, A trajectory-based analysis of coordinated team activity in a basketball game, Comput. Vis. Image Understand. 113 (5) (2009) 612–621.

[19] M. Vlachos, G. Kollios, D. Gunopulos, Discovering similar multidimensional trajectories, in: Proceedings of the International Conference on Data Engineering, 2002, pp. 673–685.

[20] J. Berclaz, F. Fleuret, P. Fua, Multi-camera tracking and atypical motion detection with behavioral maps, in: ECCV, 2008, pp. 112–125.

[21] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77(2) (1989) 257–286.

[22] J. Duchi, S. Shalev-Shwartz, T.C.Y. Singer, Efficient projections onto the ℓ_1-ball for learning in high dimensions, in: Proceedings of the International Conference on Machine Learning, Helsinki, Finland, 2008.
[23] Shun-ichi Amari, Natural gradient work efficiently in learning, Neural Comput. 10 (2) (1998) 251–276.
[24] Shun-ichi Amari, Hiroshi Nagaoka, Methods of Information Geometry, Oxford University Press, 2000.
[25] T. Boult, R. Micheals, X. Gao, M. Eckmann, Into the woods: visual surveillance of non-cooperative camouflaged targets in complex outdoor settings, Proc. IEEE 89(10) (2001) 1382–1402.

Jacinto C. Nascimento received the EE degree from Instituto Superior de Engenharia de Lisboa, in 1995, the M.Sc. and Ph.D. degrees from Instituto Superior Técnico (IST), Technical University of Lisbon, in 1998, and 2003, respectively. Currently, he is an Assistant Professor with the Informatics and Computer Engineering Department, Instituto Superior Técnico, Lisbon, and a Researcher at the Institute for Systems and Robotics. He has published over 100 publications in international journals and conference proceedings, has served on program committees of many international conferences, and has been a reviewer for several international journals. His research interests include statistical image processing, pattern recognition, machine learning, medical imaging analysis, video surveillance, general visual object classification.

Miguel Barão received the EE degree, M.Sc. and Ph.D. from Instituto Superior Técnico (IST) in 1996, 2000 and 2008, respectively. Currently, he is an Assistant Professor at the Informatics Department of Universidade de Évora, and a Researcher of INESC-ID at the Control of Dynamical Systems group. His current research interests include nonlinear and distributed control theory, information geometry, and problems at the intersection of these areas. He has been responsible or participated in several research projects on control of solar collector fields, HIV1, automotive control, video surveillance, probabilistic geometric control, among others. He is also author/coauthor of several journal and conference papers in control.

Jorge S. Marques received the E.E. and Ph.D. degrees, and the aggregation title from the Technical University of Lisbon, Portugal, in1981, 1990, and 2002, respectively. Currently, he is an Associate Professor with the Electrical and Computer Engineering Department,Instituto Superior Técnico, Lisbon, and a Researcher at the Institute for Systems and Robotics. He has published over 150 papers in international journals and conferences and he is the author of the book Pattern Recognition: Statistical and Neural Methods (IST Press, 2005, 2nd edition, in Portuguese). He was the Co-Chairman of the IAPR Conference IbPRIA 2005, President of the Portuguese Association for Pattern Recognition (2001–2003) and Associate Editor of the Statistics and Computing Journal, Springer. His research interests are in the areas of statistical image processing,shape analysis, and pattern recognition.

João M. Lemos received his Ph.D. in 1989 from IST (Instituto Superior Técnico, Technical University of Lisbon, Portugal) after extensive periods of research work at the University of Florence, Italy, and a period of experimental work at the Department of Chemical Engineering of Imperial College of Science, Technology and Medicine, London, U.K., and became an associate professor in 1997. Currently, he is a full professor (professor catedrático) of Systems Decision and Control at IST and researcher of INESC-ID, where he leads the Control of Dynamic Systems Group since its creation in 1990. He has been the Chairman of the Department of Electrical and Computer Engineering of IST and Coordinator of the post-graduation Programme in Electrical and Computer Engineering and, since 2007, he is the Chairman of the Scientific Board of INESC-ID. He co-authored 40 journal papers (ISI referenced), 10 book chapters and 180 conference papers (peer reviewed), and supervised 13 Ph.D. thesis (completed) and 27 M.Sc. dissertations. His research interests are in the area of Computer Control including adaptive and predictive control, control and estimation with multiple models, distributed control, modelling and identification and process monitoring. He has been involved, both either as responsible or participant, in projects concerning control applications to several types of industrial processes, large scale boilers in thermoelectric plants, solar collector fields and furnaces, biomedical systems (general anaesthesia, HIV 1 therapy, biochip temperature control), automotive systems and water delivery canals. He enjoys painting water colours in the open air, reading about History and fishing.