Information Geometric Algorithm for Estimating Switching Probabilities in Space-Varying HMM

Jacinto C. Nascimento, *Member, IEEE*, Miguel Barão, Jorge S. Marques, and João M. Lemos

***Abstract*— This paper proposes an iterative natural gradient algorithm to perform the optimization of switching probabilities in a space-varying hidden Markov model, in the context of human activity recognition in long-range surveillance. The proposed method is a version of the gradient method, developed under an information geometric viewpoint, where the usual Euclidean metric is replaced by a Riemannian metric on the space of transition probabilities. It is shown that the change in metric provides advantages over more traditional approaches, namely: 1) it turns the original constrained optimization into an unconstrained optimization problem; 2) the optimization behaves asymptotically as a Newton method and yields faster convergence than other methods for the same computational complexity; and 3) the natural gradient vector is an actual contravariant vector on the space of probability distributions for which an interpretation as the steepest descent direction is formally correct. Experiments on synthetic and real-world problems, focused on human activity recognition in long-range surveillance settings, show that the proposed methodology compares favorably with the state-of-the-art algorithms developed for the same purpose.**

***Index Terms*— Hidden Markov models, EM algorithm, natural gradient, parametric models, surveillance, trajectories, vector fields.**

I. Introduction

OPTIMIZATION of probability distributions is ubiquitous in science and engineering applications. Practical applications addressing this problem embrace quite diverse research areas. For instance, in route traffic optimization, aiming at selecting the route according to a given distribution, such that several routes can be used at its maximum capacity. Other application is in control of regulatory gene networks, where each network state is a specific combination or activation profile of genes. Since interactions at the molecular level are stochastic, the probabilistic approach becomes a natural formulation. Other application is in image processing, e.g. *activity classification from pedestrian's trajectories*, where the switching probability mechanism is also present; the focus on this paper. More specifically, we consider the activity recognition problem in a far-field surveillance scenario. In this problem a set of observed pedestrian trajectories are used to build a model of pedestrian's behavior in a scene. It is well known that, in many scenarios, people tend to follow a set of typical trajectories. Based on this observation, we propose to model pedestrian trajectories via a small set of vector/motion fields estimated from observed trajectory data [1], [2]. To increase the expressiveness of the model, we let each trajectory follow a probabilistic mechanism which is space-dependent; *i.e.*, the switching probability between the vector/motion fields may depend on the specific spatial location. For this purpose, we have a field of switching (stochastic) matrices on the image domain. This approach is flexible enough to represent a wide variety of trajectories and allows modeling space-varying behaviors without resorting to non-linear dynamical models, which are infamously hard to estimate from training data.

Learning of the models is performed using the Expectation-Maximization (EM) algorithm. In each EM iteration, a maximization step is performed on the vector/motion fields and on the switching matrices separately. While vector/motion fields are explicitly maximized, switching matrices do not have such explicit solutions. We therefore resort to iterative methods from the family of gradient based methods. Other methods have been tried before with success [1], [2]. Their computational cost, however, is high since a full run of a gradient method is required on each step of the EM algorithm.

To overcome the computational costs involved, this paper proposes the application of the natural gradient method. We show both theoretically and experimentally that the methodology herein proposed is tailored for estimating such transition probabilities with obvious advantages.

This paper is organized as follows: Section III presents the natural gradient algorithm and demonstrates some of the theoretical properties; Section IV presents in greater detail the space-varying HMM, more specifically, the function to be optimized; Section V shows experimental results and

Manuscript received February 6, 2014; revised July 15, 2014; accepted October 8, 2014. Date of publication October 15, 2014; date of current version October 31, 2014. This work was supported in part by the Fundação para a Ciência e a Tecnologia under Project PEst-OE/EEI/LA0009/2013, in part by the ARGUS Project under Grant PTDC/EEA-CRO/098550/2008, in part by the Instituto de Engenharia de Sistemas e Computadores-Investigação e Desenvolvimento, Lisbon, Portugal, under Grant PEst-OE/EEI/LA0021/2013, and in part by the European Union under Grant FP6-EU-IST-045062 through the Ubiquitous Networking Robotics in Urban Settings Project for the use of the real data and the UPC images. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Deniz Erdogmus.

J. C. Nascimento and J. S. Marques are with the Department of Informatics and Computer Engineering, Instituto Superior Técnico, Technical University of Lisbon, Lisbon 1049-001, Portugal, and also with the Departamento de Engenharia de Informática e de Computadores, Institute for Systems and Robotics, Lisbon 1049-001, Portugal (e-mail: jan@isr.ist.utl.pt; jsm@isr.ist.utl.pt).

M. Barão is with the Department of Informatics, Universidade de Évora, Évora 7004-516, Portugal, and also with the Control of Dynamical Systems Group, Instituto de Engenharia de Sistemas e Computadores-Investigação e Desenvolvimento, Lisbon 1000-029, Portugal (e-mail: miguel.barao@gmail.com).

J. M. Lemos is with the Systems Decision and Control Section, Instituto Superior Técnico, Technical University of Lisbon, Lisbon 1049-001, Portugal, and also with the Control of Dynamical Systems Group, Instituto de Engenharia de Sistemas e Computadores-Investigação e Desenvolvimento, Lisbon 1000-029, Portugal (e-mail: jlml@inesc-id.pt).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2363614

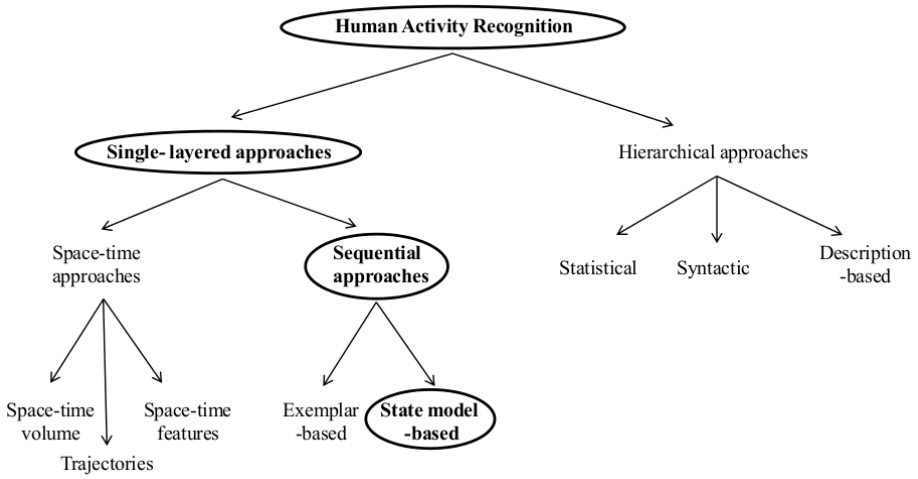


Fig. 1. Human activity taxonomy as proposed by Aggarwal and Ryoo in the review [6] where the "path" of the proposed methodology is highlighted.

compares the proposed algorithm with the state-of-the-art; finally, Section VI draws conclusions.

II. Related Work

There are several existing surveys within the area of human activity analysis [3]–[6], where quite distinct focus on this subject are available in these overviews. In this work, we describe the human activity related works following the review proposed by Aggarwal [6]. We first review the categories of the recognition methodologies describing its main characteristics, and positioning our methodology viewed by the approach-based taxonomy [6]. Then, we briefly review the works that most relate with the proposed approach.

Following the mentioned review, different class of approaches have been proposed. Basically they can be classified in *single layered* or *hierarchical* approaches. Single-layered approaches attempt to represent and recognize human activities from a sequence of images, whereas hierarchical approaches represent each human activity as being composed by several atomic (or event) actions that are more simpler to describe (see Fig. 1 for an illustration of the taxonomy). Single-layered approaches can be further divided into *space-time* and *sequential* approaches. Basically, the former considers the input video as volume, i.e. a collection of frames through time. In the latter (sequential), the human activity is considered a sequence of observations computed from the image sequence.

According to the above tree-structured taxonomy, the framework herein presented is a *single-layered* approach, since we aim at classifying human activities based on a sequence of images. Also, it is *sequential*, since the proposed approach recognizes human activities by analyzing sequence of features or observations through time converting, in this way, a sequence of images into a sequence of observations; particularly the image sequences are converted into *trajectories*. More specifically, we propose a sequential-single-layered based on a space varying HMM i.e. *state model-based* approach.

Hidden Markov Models (HMMs) have been widely proposed as suitable state model-based approach for recognizing human activities due to its effectiveness for modeling the variations of the observations in the human activities or actions. One of the first approaches is rooted in [7] that uses a standard HMM to recognize activities and [8] for recognizing human gesture using state model. Variants of HMM have also been proposed to tackle human activity analysis. One of the first examples is in [9], where a coupled HMM (CHMM) is built to model human-to-human interactions. In [10] a dynamic Bayesian Network (DBN) is used to recognize gestures of two interacting people. Another variant of the HMM for modeling human interactions is proposed in [11]. More specifically, the framework is an extension of CHMM called coupled hidden semi-Markov models (CHSMMs), where the HMM has compositional state in both space and time.

In a slightly different context, several researchers have proposed new approaches for modeling relationships and dependencies among objects and human activities to improve both object recognition and human activities performances. One such example is the approach in [12], that proposes a Bayesian method for modeling contextual relationships between four perceptual elements of human object interaction namely, object perception, reach motion, manipulation motion and object reaction. In the same line of state-model based approaches, the work in [13] demonstrates that the partial models of individual static poses can be combined with partial models of the video's motion dynamics to achieve motion classification.

All the above mentioned works act mostly at action, gesture or interaction levels, being *short-range* based approaches. Our proposal contrasts with aforementioned works in the sense that we focus on *long-range*, with a camera covering a wide area, therefore, the video resolution is low. Usually this embraces single-view, single-view/view-invariant and multi-view settings [14], [15]. Under this constraint, it is not possible to obtain a detailed feature descriptors. The typical solution is to collect only the object's positions through time along tracks, termed as *trajectories*. Still, the trajectories are a rich source of information that have been applied in other category of approaches such as in *space-time* approaches (see Fig. 1). In this category of approaches, the activity is interpreted as a set of space-time

trajectories that are capable to detail the human movements. However, the recognition process is generally regarded as a template matching process. This means that a template, learned using training trajectories, must be matched against new observed (test) trajectories to perform activity recognition. Thus, this strategy requires some similarity measure. What happens is that, the same activity may be performed in different rates, so similarity must consider these variations. To tackle this problem, different trajectory analysis problems have been addressed using pairwise (dis)similarity measures between trajectories. This includes, for instance, Euclidean [16], Hausdorff [17], [18] distances. However, when the trajectories are produced by probabilistic generative models [19]–[22], usually of the hidden Markov model (HMM) type, as we propose in this paper, this does not require alignment or registration of the trajectories being compared; moreover, it provides a solid probabilistic inference framework, based on which model parameters may be obtained from observed data.

A. Detailing the Algorithm

In our previous work [1], we have proposed a class of space-varying switched motion fields model for classifying human activities. Basically this framework is equipped with a model switching in the Markov chain, in which, the transition probabilities are estimated in a simplex probability constraint. Under this constraint, standard gradient based methods are applicable. In this paper, the methodology contrasts with [1] in sense that now, we stay away from this assumption by proposing a natural gradient method for the estimation of switching probabilities. 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.

The use of the natural gradient is motivated by the fact that the space of probability distributions is a space with its own natural metric structure that is not exploited in generic methods. This point of view is exposed in works by Amari and others [23] and is generally known as information geometry. In particular, it is suggested [24] that the natural gradient should be used in learning problems. As an application example, the natural gradient is used in [25] to solve a very high dimensional probabilistic control problem.

III. Natural Gradient

Optimization of switching probabilities is usually dealt with as a constrained optimization problem since probabilities have to lie on a probability simplex. Here, we adopt a different approach based on the information geometric framework [23], [24]. Within this framework, switching probabilities are considered as points on a differentiable manifold (a statistical manifold) and their optimization is then performed as an unconstrained optimization problem.

When dealing with categorical distributions having probabilities $(p_1, \ldots, p_K)$, one possible parameterization is to use

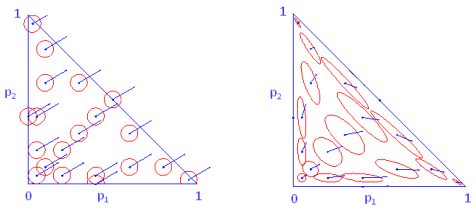


Fig. 2. Illustration of the differences between the Euclidean geometry (left) and the proposed manifold geometry (right).

the $K-1$ independent probabilities $\xi = (p_1, \ldots, p_{K-1})$ as coordinates, the remaining probability p_K being dependent and automatically computed by $p_K = 1 - \sum_{i=1}^{K-1} p_i$. This parameterization provides a global coordinate system on the manifold. Then, a metric can be introduced by defining an inner product $\langle v, w \rangle = v^T G w$ between vectors v, w belonging to the tangent space of each point ξ. A particularly interesting metric makes use of the Fisher information matrix $G = [g_{ij}]$ defined by

$$g_{ij}(\xi) = E\left[\frac{\partial \log p}{\partial \xi^i}\frac{\partial \log p}{\partial \xi^j}\right]. \tag{1}$$

This metric has the property that it is invariant with respect to coordinate transformations and is central to the definition of natural gradient. The invariance property implies that the geometry does not depend on a particular parametrization for the probability distribution and justifies the name *natural*.

The gradient vector of a function f is the (contravariant) vector ∇f such that $\langle \nabla f, v \rangle = \mathrm{d}f(v)$ holds for any vector v. In this equation, $\mathrm{d}f$ denotes the differential of the function f (a 1-form whose components are given by the partial derivatives of f). In the matrix convention adopted henceforth, the differential is written as the row matrix with components $\mathrm{d}f = \left[\frac{\partial f}{\partial p_1} \quad \cdots \quad \frac{\partial f}{\partial p_{k-1}}\right]$ and the gradient vector as a column matrix. Specializing the gradient vector computation for the statistical manifold of categorical probability distributions, and using the Fisher information metric (1) above, yields

$$\nabla f = G^{-1}(\mathrm{d}f)^T = \xi \circ (\mathrm{d}f)^T - \xi\,(\mathrm{d}f \cdot \xi), \tag{2}$$

where the operator $\circ$ denotes the Hadamard product (elementwise product of vectors) and $\cdot$ is the usual dot product. Using the formula on the right hand side, the gradient can be computed avoiding the explicit construction of the Fisher information matrix and its inverse, yielding linear time computational complexity and not requiring additional memory. Also, it can be shown [25] that ∇f vanishes on the boundaries of the probability simplex. This latter fact alone, turns probability optimization into an unconstrained problem, since it is not possible to move out of the simplex using an appropriately bounded, but positive, optimization step.

To illustrate the effect of the new geometry culminating in the transformation (2), Fig. 2 shows how arbitrary differentials $\mathrm{d}f$ are transformed into natural gradient vectors ∇f. While the differential $\mathrm{d}f$ can arbitrarily point in any direction (left image), depending only on the function f, its corresponding gradient vector also depends on the metric G and is thus affected by the point on the manifold where the gradient is

computed (right image). For instance, near a simplex border, the metric tends to loose the corresponding dimension, which can be seen by the ellipsis becoming narrower.

The maximization of a generic function f is then performed using the natural gradient method

$$\xi_{t+1} = \xi_t + \eta \nabla f, \tag{3}$$

where the scalar η is a positive step size. This equation is iterated until convergence is attained. To ensure that the updated probabilities correspond to a valid probability distribution, the step size has to satisfy the bounds

$$0 < \eta < \frac{1}{\mathrm{d}f \cdot \xi - \alpha}, \qquad \alpha = \min\{0, (\mathrm{d}f)_1, \ldots, (\mathrm{d}f)_{K-1}\}. \tag{4}$$

Although (3) has the appearance of a standard gradient method, the fact that we are using the natural gradient, derived from the Fisher information metric, provides additional advantages in some classes of problems. A particular example is when the objective function f is the Kullback-Leibler divergence between two probability distributions, one of them being the probability distribution under optimization. Since the Hessian of the Kullback-Leibler divergence $D(p_{\xi^*} \| p_{\xi^*+\Delta\xi})$ is precisely the Fisher information matrix G defined in (1), the gradient method (3), rewritten as

$$\xi_{t+1} = \xi_t + \eta G^{-1} (\mathrm{d}f)^T, \tag{5}$$

shows that for small $\Delta\xi$ (i.e., near the optimal point ξ^*) it can be interpreted as a Newton method, thus providing quadratic asymptotic convergence. This fact will be exploited later to justify the observed gains in convergence speed of the surveillance problem.

IV. Space-Varying HMM

In this section, we illustrate how the natural gradient methodology described in Section III can be useful for estimating the parameters in a generative model as detailed in Section IV-A, from which, the trajectories are drawn. The Markov chain herein formulated is space varying, and it will be shown that the natural gradient can accurately estimate the space varying switching probabilities in the chain. The application used focus on the trajectory classification for human activity recognition, one of the core research topics in surveillance systems as we illustrate in Section V.

A. Generative Motion Model - Multiple Vector Fields

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 [1]

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{T}_{k_t}(\mathbf{x}_{t-1}) + \mathbf{w}_t, \qquad t = 2, \ldots, L, \tag{6}$$

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 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 $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_L\}$ and its underlying sequence of active fields $\mathbf{k}$, under the model parameters $\boldsymbol{\Theta} = (\mathcal{T}, \mathbf{B}, \boldsymbol{\Sigma})$, is given by

$$p(\mathbf{x}, \mathbf{k} | \boldsymbol{\Theta}) = p(\mathbf{x}_1) P(k_1) \prod_{t=2}^{L} p(\mathbf{x}_t | \mathbf{x}_{t-1}, k_t) p(k_t | k_{t-1}, \mathbf{x}_{t-1}). \tag{7}$$

From (7), 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 [1] 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

$$\mathbf{T}_k(\mathbf{x}) = \sum_{n=1}^{N} \mathbf{t}_k^{(n)} \phi_n(\mathbf{x}), \qquad \mathbf{B}(\mathbf{x}) = \sum_{n=1}^{N} \mathbf{b}^{(n)} \phi_n(\mathbf{x}) \tag{8}$$

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

B. Learning the Generative Model

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

$$\begin{aligned} \widehat{\boldsymbol{\Theta}} &= \arg\max_{\boldsymbol{\Theta}} \sum_{\mathcal{K}} p(\mathcal{X}, \mathcal{K} | \boldsymbol{\Theta})\, p(\boldsymbol{\Theta}) \\ &= \arg\max_{\boldsymbol{\Theta}} \sum_{\mathcal{K}} \prod_{s=1}^{S} p(\mathbf{x}^{(s)}, \mathbf{k}^{(s)} | \boldsymbol{\Theta})\, p(\boldsymbol{\Theta}) \end{aligned} \tag{9}$$

where each factor $p(\mathbf{x}^{(s)}, \mathbf{k}^{(s)} | \boldsymbol{\Theta})$ has the form given in (7), 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.

C. The Complete Log-Likelihood

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

$$\begin{aligned} &Q(\boldsymbol{\Theta};\widehat{\boldsymbol{\Theta}}) \\ &\quad\equiv \mathbb{E}_{\mathcal{K}}\left[\log p(\mathcal{X},\mathcal{K}|\boldsymbol{\Theta})|\mathcal{X},\widehat{\boldsymbol{\Theta}}\right] \\ &\quad= \underbrace{\sum_{s=1}^{S}\sum_{t=2}^{L_s}\sum_{l=1}^{K}\bar{w}_{t,l}^{(s)}\log\mathcal{N}(\mathbf{x}_t^{(s)}|\mathbf{x}_{t-1}^{(s)}+\mathbf{T}_l(\mathbf{x}_{t-1}^{(s)}),\sigma_l^2\mathbf{I})}_{\widetilde{\mathcal{A}}(\boldsymbol{\Theta};\widehat{\boldsymbol{\Theta}})} \\ &\qquad+\underbrace{\sum_{s=1}^{S}\sum_{t=2}^{L_s}\sum_{l=1}^{K}\sum_{g=1}^{K}\bar{w}_{t,g,l}^{(s)}\log B_{g,l}(\mathbf{x}_{t-1}^{(s)})}_{\widetilde{\mathcal{B}}(\boldsymbol{\Theta};\widehat{\boldsymbol{\Theta}})}, \end{aligned} \tag{10}$$

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

The M-step maximizes the Q-function in (10) 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 $\widetilde{\mathcal{A}}(\boldsymbol{\Theta};\widehat{\boldsymbol{\Theta}})$ in (10)) can be achieved following the same strategy as in [1].

To estimate the term $\widetilde{\mathcal{B}}(\boldsymbol{\Theta};\widehat{\boldsymbol{\Theta}})$ in (10) is much more difficult, and we delve into the framework as described in Section III.

The term $\widetilde{\mathcal{B}}(\boldsymbol{\Theta};\widehat{\boldsymbol{\Theta}})$ is the one which allows the estimation of the transition matrix and it is the most important in many human activity recognition problems. In fact, it is a very specific feature that depends on the class itself, i.e. the switching probabilities among the motion regimes can contribute as a discriminative information for the recognition task, as will be discussed further (see Section V).

We now argue that the use of the natural gradient method to estimate $B_{g,l}(\mathbf{x}_{t-1}^{(s)})$ has the potential to obtain a higher convergence rate than standard gradient methods. To see this, first notice that $\widetilde{\mathcal{B}}(\boldsymbol{\Theta};\widehat{\boldsymbol{\Theta}})$ in (10) is the term that we are actually interested in, when maximizing $Q(\boldsymbol{\Theta};\widehat{\boldsymbol{\Theta}})$ with respect to the transition probabilities. This expression can be rewritten as follows:

$$\mathcal{F}(B(\mathbf{x}_{t-1})) = \sum_{s=1}^{S}\sum_{t=2}^{L_s}\sum_{l=1}^{K}\sum_{g=1}^{K}\bar{w}_{t,g,l}^{(s)}\log B_{g,l}(\mathbf{x}_{t-1}^{(s)}). \tag{11}$$

Factorizing the joint distribution $\bar{w}_{t,g,l}^{(s)} = \bar{w}_{t,g}^{(s)}\bar{w}_{t,g\to l}^{(s)}$, using $\bar{w}_{t,g\to l}^{(s)} = P[w_{t,g\to l}^{(s)} = 1|w_{t-1,g}^{(s)} = 1,\mathbf{x}^{(s)},\widehat{\boldsymbol{\Theta}}]$, yields

$$\begin{aligned} &\mathcal{F}(B(\mathbf{x}_{t-1})) \\ &= \sum_{s=1}^{S}\sum_{t=2}^{L_s}\sum_{g=1}^{K}\bar{w}_{t,g}^{(s)}\sum_{l=1}^{K}\bar{w}_{t,g\to l}^{(s)}\log B_{g,l}(\mathbf{x}_{t-1}^{(s)}) \\ &= \sum_{s=1}^{S}\sum_{t=2}^{L_s}\sum_{g=1}^{K}\bar{w}_{t,g}^{(s)}\sum_{l=1}^{K}\bar{w}_{t,g\to l}^{(s)}\log\frac{B_{g,l}(\mathbf{x}_{t-1}^{(s)})\bar{w}_{t,g\to l}^{(s)}}{\bar{w}_{t,g\to l}^{(s)}} \\ &= \sum_{s=1}^{S}\sum_{t=2}^{L_s}\sum_{g=1}^{K}\bar{w}_{t,g}^{(s)}\left(-D\big(\bar{w}_{t,g\to l}^{(s)}\|B_{g,l}(\mathbf{x}_{t-1}^{(s)})\big) - H(\bar{w}_{t,g\to l}^{(s)})\right), \end{aligned} \tag{12}$$

where $D(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence between two probability distributions and $H(\cdot)$ denotes the entropy function. When maximizing with respect to the transition probabilities, only the Kullback-Leibler divergence term is of interest. We can now provide an interpretation of the M-step as the minimization of the expected Kullback-Leibler divergence between $\bar{w}_{t,g\to l}^{(s)}$ and switching probabilities $B_{g,l}(\mathbf{x}_{t-1}^{(s)})$, *i.e.*

$$\min\sum_{s=1}^{S}\sum_{t=2}^{L_s}\sum_{g=1}^{K}\bar{w}_{t,g}^{(s)}\,D\big(\bar{w}_{t,g\to l}^{(s)}\parallel B_{g,l}(\mathbf{x}_{t-1}^{(s)})\big). \tag{13}$$

This expression can be interpreted as follows. For sufficiently fine grained grid, the transition probabilities $B_{g,l}(\mathbf{x}_{t-1}^{(s)})$ can exactly match the weights $\bar{w}_{t,g\to l}^{(s)}$ and the Kullback-Leibler divergence attains its minimum value of zero. For coarser grids, this is no longer the case, and we are minimizing the discrepancy in switching probabilities weigthed by probability of the previous active field (if a field has zero probability, the transition probabilities from that field to any other are undefined). See additional comments in the Appendix.

V. Experimental Results

In this section we present results concerning the convergence of the proposed algorithm and trajectory (activities) classification using several synthetic examples. Also, we present results in a real scenario illustrating the performance of the proposed technique for classification of the pedestrians activities (from trajectories) typically found in far-field surveillance scenarios. In the experimental evaluation presented, we also perform a comparison between the proposed method described in Section IV-B and the gradient projection (GP) algorithm [26] used in [1]. The two main components of the GP algorithm are: (i) the computation of the gradient of the objective function and the projection onto the constraint set (i.e. stochastic matrices). Concerning the latter (i.e. the projection), it consists in projecting each row of the transition matrix onto a probability simplex, for which a recent approach has been proposed (see [27] for an in depth review).

Throughout the section of the experimental evaluation, we assume that we have a known number A of different activity classes (type of trajectories), i.e. $a \in \{1,\ldots,A\}$, and that we have a subset of trajectories from each of these activity classes, $\mathcal{X}^1,\ldots,\mathcal{X}^A$. We will denote the set of the fields and parameters corresponding to each activity class a as $\boldsymbol{\Theta}^a = (\mathcal{T}^a,\mathbf{B}^a,\boldsymbol{\Sigma}^a)$, for $a = 1,\ldots,A$. In some cases, one or more of these collections of parameters may be shared among the classes; for example, if the motion fields are common among the classes and only the switching matrices differ, we have $\mathcal{T}^a = \mathcal{T}$ and $\boldsymbol{\Sigma}^a = \boldsymbol{\Sigma}$, for $a = 1,\ldots,A$.

A. Convergence of the Natural Gradient

1) Example 1: This example illustrates the convergence of the proposed technique at estimating the space-varying stochastic matrix (see (10)) in the context of a synthetic trajectory (activities) example.

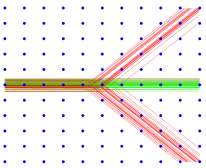


Fig. 3. Two synthetic trajectory classes. Straight (blue) and disperse (red) trajectories containing different motion models. Grid nodes ($n = 121$) are shown in blue dots.

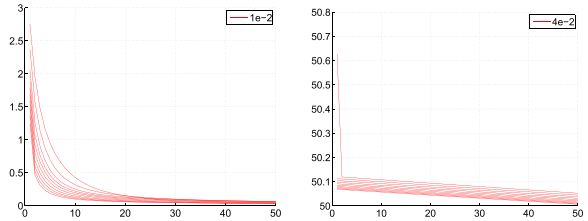


Fig. 4. Convergence of the gradient during the 10 iterations of the EM algorithm. (a) results using the proposed methodology ($\eta = 1 \times 10^{-2}$) (b) results using the GP method used in [1] ($\eta = 4 \times 10^{-2}$).

Fig. 3 illustrates two classes (activities) of trajectories. One class contains one left-right (horizontal) motion model to describe, say, the *straight* activity. This activity has zero probability of switching, i.e. identity transition matrices everywhere in the grid.[1] The second activity contains three motion models, one left-right horizonal motion as in the previous class, and two diagonal motions to describe the *disperse* activity. In this case, the trajectories can turn up or down. Thus, to generate this class, the entries that allow to switch from the horizontal motion model to one of the diagonal motion models are set to 0.5. The remaining entries follow an identity structure. All the trajectories start roughly at the image point $[0, 0.5]^T$ of the unit square as illustrated in the figure.

To illustrate the convergence velocity of the proposed technique, we turn the experiment as simple as possible: we assume that we know beforehand the velocity fields $\mathcal{T}$ (i.e. the first term in (10) is known). Under this assumption, we only have to estimate, the weights $\bar{w}_{t,g,l}^{(s)}$ and the transition matrix $B_{g,l}(\mathbf{x}_{t-1}^{(s)})$ for all the trajectories. In this experiment we used 10 iterations for the EM algorithm and 50 internal iterations in the M-step to estimate the transition matrix B.[2] Fig. 4 illustrates the convergence velocity for the best step-size η found for each methodology. Each descendent line has 50 points, corresponding to the number of iterations used in the M-step. It is illustrated a total of 10 lines corresponding to each of the EM iterations. We observe that the natural gradient behaves a quasi-Newton, variable metric, method as it approaches the optimum. This justification is due to the Fisher information metric matrix that plays a fundamental role in the framework. It is seen the quadratic behavior of the proposed methodology vs. the linear behavior of the project simplex

[1] In all the experiments presented a grid of 11×11 nodes is used to represent the vector fields and the field of stochastic matrices.

[2] Although 50 iterations were used for the illustration purposes, only 10 iterations suffice as illustrated later in this section.

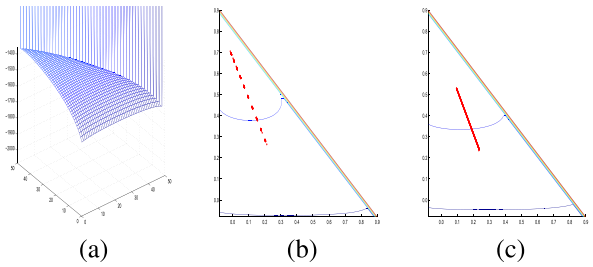


Fig. 5. Gradient direction during the convergence of the EM. (a) 3D surface of the cost function in (10); (b) gradient direction in the proposed approach and (c) in the GP algorithm. The level curves of the cost function are in blue; the trajectories performed in the grid node are in red.

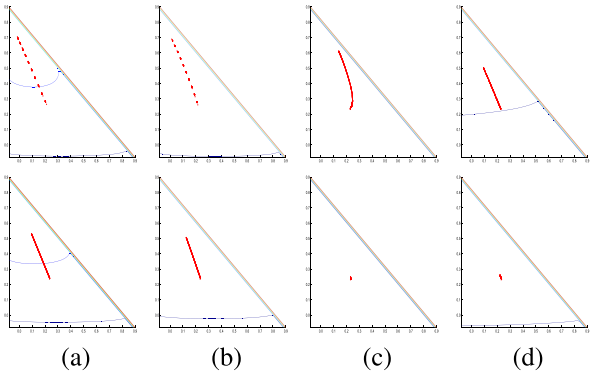


Fig. 6. Gradient direction during the convergence of the EM. The same node grids are shown for the natural gradient (top) and GP algorithm (bottom). From left to right columns: $56th$, $60th$, $62th$ and $69th$ grid nodes. The level curves are in blue; the grid node trajectories are in red.

used in [1]. Also we observe that the convergence is very fast for the method proposed herein.

Fig. 5(a) shows the 3D surface of cost the function to be minimized for the natural gradient and for the GP algorithm. Fig. 5 (b,c) shows the gradient direction during the convergence process along with the level curves of the cost function (i.e. second term of $Q(\boldsymbol{\Theta}; \widehat{\boldsymbol{\Theta}})$ in (10)). In these two images we illustrate the trajectories performed in the $56th$ node in the grid (total of 121 nodes) for each algorithm. The level curves (contour of the cost function) are shown as a function of the entries (b_{11}, b_{12}) of the transition matrix. In this experiment, the transition matrices of all the nodes are equally initialized with entries $b_{i,j} = 0.33$, (with $i, j = 1, \ldots, 3$). This can be seen as the starting point of all the trajectories depicted in Fig. 5 (b,c).

Fig. 6 shows more examples of the direction of the gradient in both methodologies.[3] It can be seen that the natural gradient does not seem to intersect the level curves orthogonally. This observation is misleading since these figures are designed in the Euclidean setting, while the computations are done with respect to the Fisher metric. The orthogonality is in fact enforced in the correct metric. Figure 6(c) also shows the effect of the metric on the trajectory when probabilities get close to the simplex boundary. While in Euclidean geometry a gradient

[3] We do not show the 3D surface of the cost function since it is similar the surfaces shown in Fig. 5.

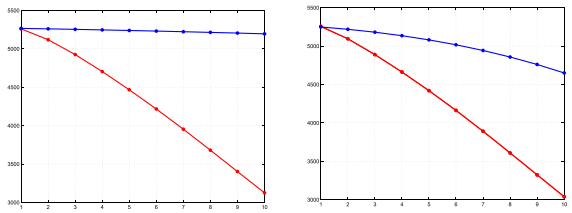


Fig. 7. Decrease of the cost function using the natural gradient (red line) and the GP method (blue line). (left) convergence with 10 iterations; (right) convergence with 50 iterations.

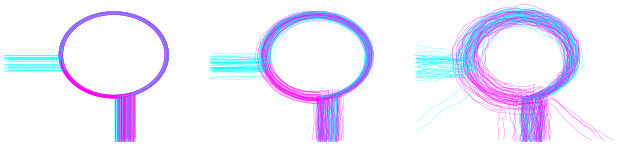


Fig. 8. Two different type of trajectories with similar switching. One class is circular, having its entry and exit points at the same bottom region of the image (magenta lines). The other class (cyan color) also performs the same rotation but only in three quarters, having the same entry point as above (at the bottom) and the exit point at the left region of the image.

flow could go directly across the simplex boundary, that does not happen in the geometry use here.

Fig. 7 shows the decrease of the cost function $Q(\boldsymbol{\Theta}\,;\,\widehat{\boldsymbol{\Theta}})$ as a function of the EM iterations; these results were obtained for 10 iterations in the M-step. This procedure is the same for both methods. In this experiment we introduce an additional experiment that uses only 10 iteration in the M-step. This is to illustrate the variation of the decrease varying the number of internal iterations. Again the superiority of the proposed method is evident exhibiting faster convergence in both cases. Please note the scale in Fig. 4.

B. Trajectory Classification

1) Example 2: In this example we present two trajectory classes shown in Fig. 8. The two activities are similar. One is characterized by circular trajectories with entry and exit points at the same bottom region of the image (see magenta lines in Fig. 8(a)). The other (cyan color) also performs a rotation but the duration angle is smaller (three quarters) having the same entry point as above (at the bottom) and the exit point at the left region of the image. In this example the motion presented in both classes overlap in a quite significant image region and it is performed in a counterclockwise direction.

In this experiment we assumed the number of motion models previously known, thus we set $K = 2$. Nevertheless, this could be automatically determined by using a discriminative based approach as proposed in [28]. As above mentioned, we perform a comparison between the proposed method and the gradient projection (GP) [1], [26].

To perform the comparison, 10 experiments were conducted. For each experiment, we generated 100 (see (6)) training and 100 testing trajectories. For the training trajectories we set $\sigma^2_{\text{trn}} = 10^{-4}$, for the testing set we progressively increased the values of the dynamic noise in the following range:

$$\mathcal{R}_{\sigma^2_{\text{tst}}} = \{\sigma^2_{\text{trn}}, 2\sigma^2_{\text{trn}}, 5\sigma^2_{\text{trn}}, 8\sigma^2_{\text{trn}}, 10\sigma^2_{\text{trn}}, 16\sigma^2_{\text{trn}}, 20\sigma^2_{\text{trn}}, 32\sigma^2_{\text{trn}}, 50\sigma^2_{\text{trn}}, 100\sigma^2_{\text{trn}}\}$$

TABLE I
PARAMETERS INITIALIZATION FOR BOTH METHODOLOGIES

Parameter		Initial conditions
$\Sigma = \{\sigma^2_{(1)}, .., \sigma^2_{(N)}\}$	$\rightarrow$	1×10^{-3}
$\mathcal{T} = \{\mathbf{t}_{(1)}, .., \mathbf{t}_{(N)}\}$	$\rightarrow$	random in $[-0.01 \quad 0.01]$
$\mathcal{B} = \{\mathbf{b}_{(1)}, .., \mathbf{b}_{(N)}\}$	$\rightarrow$	$\begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$

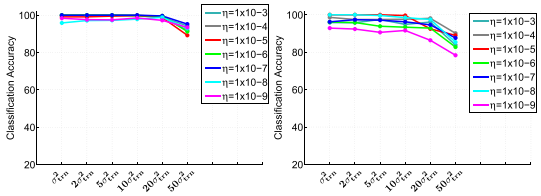


Fig. 9. Selection of the η parameter for the natural gradient method (left) and for the GP algorithm (right).

(each experiment contains 100 test trajectories with σ^2_{tst} set to a given value in $\mathcal{R}_{\sigma^2_{\text{tst}}}$). This strategy permits to verify the robustness of the proposed approach against trajectory mismatch. The performance is measured in terms of classification accuracy which is accomplished by simply using the forward E-step.

To perform a fair comparison, the initial conditions of the EM algorithm are the same for both of the methodologies. The parameter values are set as follows: we used 7 iterations for the EM, $K = 2$ motion fields, and 10 iterations in the M-step to estimate the transition matrices in all nodes of the grid (contained in $\mathbf{B}$). Due to the EM dependence regarding the initialization, we performed 8 runs of the EM to achieve significance in the performance statistics. The remaining parameter initialization follows configuration depicted in Table I.

Before presenting the accuracy performance, the first step to be accomplished is that of finding the step-size η. This is determined by cross validation in the interval $\mathcal{R}_\eta = \{10^{-3}, \ldots, 10^{-9}\}$[4] and (as above) using test trajectories with dynamic noise in the interval range of $\mathcal{R}_{\sigma^2_{\text{tst}}} = \{\sigma^2_{\text{trn}}, 2\sigma^2_{\text{trn}}, 5\sigma^2_{\text{trn}}, 10\sigma^2_{\text{trn}}, 20\sigma^2_{\text{trn}}, 50\sigma^2_{\text{trn}}\}$.

Fig. 9 shows the performance when the step size varies. We see that the proposed framework exhibits more stable results. For the GP algorithm, the results start to degrade for $\eta < 1 \times 10^{-9}$. Also, the natural gradient achieves higher classification accuracy for higher values of the dynamical noise σ^2_{tst} presented in the test sequences, and smaller variance in the range interval of this parameter.

Fig. 10 shows detailed statistical results for the best values of the step size in the range $\mathcal{R}_\eta$. We can see the superiority of the proposed framework, specially for higher values of the dynamical noise.

C. Real Data

We now consider the application of the proposed algorithm in a real setting. The images were obtained from a remote and

[4]The best results for both of the methodologies were found in this range. Outside this range we observed a decreasing performance for higher values of σ^2_{tst}.

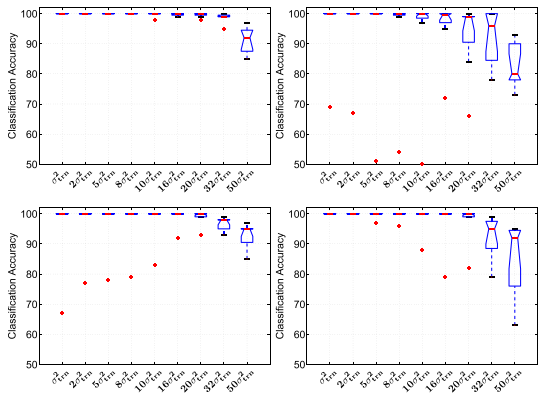


Fig. 10. Performance of the natural gradient (left column) and GP algorithm (right column) for two values of the step size η: $\eta = 1 \times 10^{-6}$ (top), 10^{-7}, $\eta = 1 \times 10^{-8}$ (bottom).

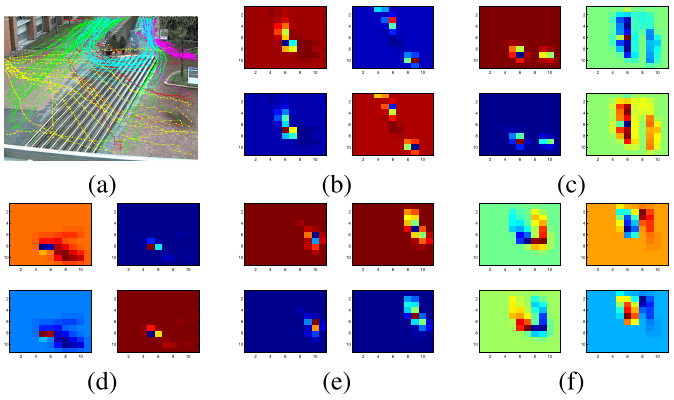


Fig. 11. (a) Trajectories from most common activity classes performed by the students in the Universitat Politécnica de Catalunya: walking and stepping up the stairs (red), walking along (green), crossing and stepping up the stairs (yellow), pass diagonally up (magenta) and turning the Campus (cyan). Switching matrices estimated using rhe natural gradient for the following activities: (b) walking and stepping up the stairs (c) walking along (d) crossing and stepping up the stairs (e) pass diagonally up and (f) turning the campus.

fixed network camera located at the campus of the Universitat Politécnica de Catalunya (UPC) Barcelona.[5] The camera was continuously streaming during several hours. Several classes of trajectories were observed and thus considered for classification using the proposed approach. The trajectories were obtained by tracking the pedestrians using the Lehigh Omnidirectional Tracking Systems (LOTS) algorithm [29]. Fig. 11(a) depicts some of the class-trajectories taking place at the above scenario (each color denotes a different activity class). The activity-classes can be framed as follows: (a_1) *walking and stepping up the stairs* (from left to right direction); (a_2) *walking along* (up motion); (a_3) *crossing and stepping up the stairs* (motion from bottom-right to up-left region); (a_4) *pass diagonally up* (from right to left); (a_5) *turning the Campus* (counterclockwise direction).

Recall that this is a challenging example since the motions (i.e. vector fields) are quite similar among classes. Only the transitions may contain specific information regarding each class of trajectories.

[5]The data was acquired in the context of the European Project FP6-EU-IST-045062, URUS - Ubiquitous Networking Robotics in Urban Settings.

An illustrative example is shown in Fig. 11 where we estimate the transition matrices for each trajectory class enrolled in the scenario.

To perform the experimental evaluation, the following issues are taken into consideration:

- *Number of motion fields*. Contrasting with the synthetic example, where we assumed the number of fields to be known, here, the model order has to be automatically discovered. To accomplish this we vary the number of motion models in the interval $K \in \{1, \ldots, 6\}$.
- *Initializations of the EM*. In order to improve the results we used eight different initializations for the EM method (i.e. eight runs of the EM).
- *Cross Validation* (CV). Since the number of trajectories is finite, we perform a 5-fold cross validation. The splitting between the training and test sets is random, but guaranteeing a balanced set of trajectories for each class.
- *Range of the step size* η. As in the synthetic example, we considered the range of this parameter $\eta \in \{1 \times 10^{-3}, \ldots, 1 \times 10^{-9}\}$ in which both approaches exhibit higher accuracy in the trajectory classification.

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 $\mathcal{F} \in \{1, \ldots, 5\}$. 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 $\mathbf{\Theta}_K^{(a)} = (\mathcal{T}, \mathbf{B}^{(a)}, \mathbf{\Sigma})$, for $a \in \{1, \ldots, A\}$, where only the switching matrices differ among the classes, and K is the number of (shared) vector fields.

Table II presents the running times of the two approaches. In this experiment, we set the value of the step-size for each algorithm. We vary the number of motion models K as shown in the table. The obtained results report the mean running time spent in seconds over the eight EM initializations for one fold (the results for the remaining folds do not change). These running time figures are obtained for the M-step when updating only the transition matrix for a single internal iteration (as in the synthetic case, 10 M-step iterations are also used). It is important to mention that these scores are obtained with an unoptimized Matlab implementations on a computer with a Intel Core i5 and 4GB of RAM.

From the Table II it can be seen that our claims stated earlier in the paper are confirmed: the natural gradient behaves asymptotically as a Newton method and yields faster convergence.

Fig. 12 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. We do not show the results for $K = 1$ since both methodologies provide lower performance accuracy. From this figure, we conclude that both approaches are remarkably competitive providing high accuracy rates. Notice however, that the natural gradient exhibits better performance at lower

TABLE II
COMPARISON OF THE RUNNING TIMES FIGURES (MEAN AND STANDARD DEVIATION IN SECS.) OF THE APPROACHES

	N° of motion fields				
	2	3	4	5	6
Gradient Projection	0.22(0.04)	0.22(0.03)	0.23(0.02)	0.24(0.01)	0.24(0.002)
Natural Gradient	0.05(0.01)	0.05(0.01)	0.05(0.01)	0.05(0.00)	0.05(0.00)

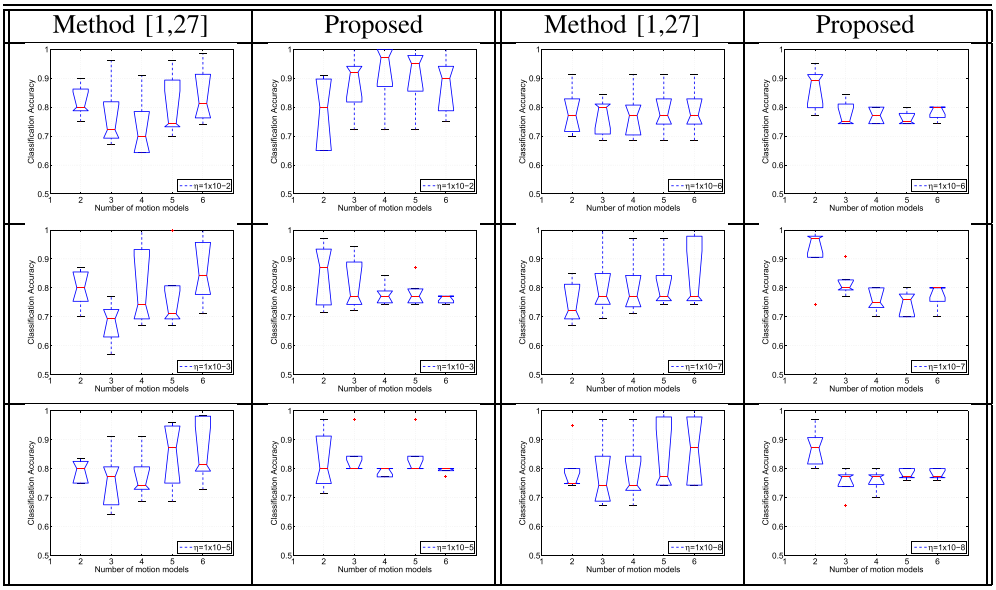


Fig. 12. Comparison of the two methodologies varying the step size η and the number of motion fields K.

number of K (lower complexity). It is seen that the natural gradient provides higher accuracy comparing to [1], using a much smaller number of motion models providing less complexity. Fig. 12 shows that the higher accuracies occur at $K = 2$, while in the GP algorithm the higher accuracies are obtained for higher model orders, i.e. $K \in \{5, 6\}$. Also, the obtained variance is smaller no matter the model order used K. This suggests that the algorithm deals well when sharing the motion fields among different type of trajectories classes not jeopardizing the classification accuracy.

VI. CONCLUSIONS

This paper presented an innovative approach to the estimation of space varying switching probabilities in the context of human activity recognition. The proposed algorithm compares favorably with sate of the art methods applied to the same problem. Furthermore, the proposed algorithm has the advantage that the optimization is performed in the manifold of probability distributions using the natural gradient with respect to the Fisher information metric in an unconstrained setup. This allows a significant reduction on the computational complexity of previous constrained optimization methods. The proposed methodology was tested and validated both on synthetic and real data. The latter obtained from surveillance videos. It is shown that the natural gradient method converges faster and attained better accuracy for the same computational effort.

APPENDIX

Given a set of trajectories represented as sequences of points, it is possible to define a sufficiently fine grained grid so that each grid's square contains either one or zero points of the trajectories. With this grid, on could assign the weights $\bar{w}_{t,g\to l}^{(s)}$ to the transition matrices $\mathbf{b}^{(n)}$ so that the Kullback-Leibler divergence attains its minimum value of zero:

$$\min \sum_{s=1}^{S}\sum_{t=2}^{L_s}\sum_{g=1}^{K} \bar{w}_{t,g}^{(s)}\, D\big(\bar{w}_{t,g\to l}^{(s)} \parallel B_{g,l}(\mathbf{x}_{t-1}^{(s)})\big) = 0. \quad (14)$$

For coarser grids, this is no longer possible in general and the minimization is performed instead by the natural gradient method. To analyze the asymptotic behavior of the algorithm we compute the Hessian of the cost function near the optimum.

The cost is a function of the K^2N transition probabilities $\mathbf{b}_{g,l}^{(n)}$, which we vectorize into a single vector ξ_i, each i indexing a particular (g, l, n) combination. The Hessian of (11) then yields

$$\frac{\partial^2 \mathcal{F}}{\partial \xi_i \partial \xi_j} = \sum_{s=1}^{S}\sum_{t=2}^{L_s}\sum_{g=1}^{K} \bar{w}_{t,g}^{(s)} \underbrace{\left(-\sum_{l=1}^{K} \bar{w}_{t,g\to l}^{(s)} \frac{\partial^2 \log B_{g,l}(\mathbf{x}_{t-1}^{(s)})}{\partial \xi_i \partial \xi_j}\right)}_{\tilde{G}_{i,j}(s,t,g)}. \quad (15)$$

It can be seen that $\tilde{G}_{i,j}(s, t, g)$ resembles the definition of Fisher Information, with the difference that it depends on two probability distributions $\bar{w}_{t,g\to l}^{(s)}$ and $B_{g,l}(\mathbf{x}_{t-1}^{(s)})$ instead of a single one. If these distributions are close, then $\tilde{G}$ is closer to the Fisher Information matrix (FIM), and the Hessian can be interpreted for each time and trajectory (t, s) as the expected FIM weighted by the probabilties $\bar{w}_{t,g}^{(s)}$.

The fact that the natural gradient method uses the inverse FIM to compute the natural gradient, as shown in (5) of Section III, allows us to expect a performance close to that of Newton method (which uses the inverse Hessian instead). This performance was confirmed experimentally, but theoretical guarantees do not seem simple to obtain for this particular problem structure, and are still an open issue.

REFERENCES

[1] J. C. Nascimento, M. A. T. Figueiredo, and J. S. Marques, "Activity recognition using mixture of vector fields," *IEEE Trans. Imag. Process.*, vol. 22, no. 5, pp. 1712–1725, May 2013.

[2] J. C. Nascimento, J. S. Marques, and J. M. Lemos, "Modeling and classifying human activities from trajectories using a class of space-variant parametric motion fields," *IEEE Trans. Imag. Process.*, vol. 22, no. 5, pp. 2066–2080, May 2013.

[3] D. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan, "Computational studies of human motion: Part 1, tracking and motion synthesis," *Found. Trends Comput. Graph. Vis.*, vol. 1, nos. 2–3, pp. 77–254, 2005.

[4] R. Poppe, "Vision-based human motion analysis: An overview," *Comput. Vis. Imag. Understand.*, vol. 108, nos. 1–2, pp. 4–18, Oct. 2007.

[5] V. Krüger, D. Kragic, A. Ude, and C. Geib, "The meaning of action: A review on action recognition and mapping," *Adv. Robot.*, vol. 21, no. 13, pp. 1473–1501, 2007.

[6] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, p. 16, Apr. 2011.

[7] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 1992, pp. 379–385.

[8] A. F. Bobick and A. D. Wilson, "A state-based approach to the representation and recognition of gesture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 12, pp. 1325–1337, Dec. 1997.

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

[10] S. Park and J. Aggarwal, "A hierarchical Bayesian network for event recognition of human actions and interactions," *Multimedia Syst.*, vol. 10, no. 2, pp. 164–179, Aug. 2004.

[11] P. Natarajan and R. Nedvatia, "Coupled hidden semi-Markov models for activity recognition," in *Proc. IEEE Workshop Motion Video Comput. (WMVC)*, Feb. 2007, p. 10.

[12] A. Gupta and L. S. Davis, "Objects in action: An approach for combining action understanding and object perception," in *Proc. IEEE Conf. CVPR*, Jun. 2007, pp. 1–8.

[13] R. Filipovych and E. Ribeiro, "Learning human motion models from unsegmented videos," in *Proc. IEEE Conf. CVPR*, Jun. 2008, pp. 1–7.

[14] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

[15] X. Ji and H. Liu, "Advances in view-invariant human motion analysis: A Review," *IEEE Trans. Syst., Man, Cybern., C, Appl. Rev.*, vol. 40, no. 1, pp. 13–24, Jan. 2010.

[16] Z. Fu, W. Hu, and T. Tan, "Similarity based vehicle trajectory clustering and anomaly detection," in *Proc. IEEE ICIP*, vol. 1. Sep. 2005, pp. II-602–II-605.

[17] I. N. Junejo, O. Javed, and M. Shah, "Multi feature path modeling for video surveillance," in *Proc. 17th Int. ICPR*, vol. 2. Aug. 2004, pp. 716–719.

[18] X. Wang, K. Tieu, and E. Grimson, "Learning semantic scene models by trajectory analysis," in *Proc. ECCV*, vol. 3. 2006, pp. 110–123.

[19] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh, "Activity recognition and abnormality detection with the switching hidden semi-Markov model," in *Proc. IEEE Comput. Soc. Conf. CVPR*, vol. 1. Jun. 2005, pp. 838–845.

[20] J. Nascimento, M. A. T. Figueiredo, and J. S. Marques, "Independent increment processes for human motion recognition," *Comput. Vis. Imag. Understand.*, vol. 109, no. 2, pp. 126–138, Feb. 2008.

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

[22] X. Wang, K. T. Ma, G.-W. Ng, and W. E. L. Grimson, "Trajectory analysis and semantic region modeling using a nonparametric Bayesian model," in *Proc. CVPR*, Jun. 2008, pp. 1–8.

[23] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry*. London, U.K.: Oxford Univ. Press, 2000.

[24] S.-I. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, no. 2, pp. 251–276, 1998.

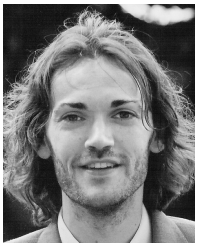
[25] M. Barão and J. M. Lemos, "An efficient Kullback–Leibler optimization algorithm for probabilistic control design," in *Proc. Medit. Conf. Control Autom.*, Jun. 2008, pp. 198–203.

[26] J. Nocedal and S. Wright, *Numerical Optimization*. New York, NY, USA: Springer-Verlag, 2006.

[27] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the ℓ_1-ball for learning in high dimensions," in *Proc. ICML*, 2008, pp. 272–279.

[28] J. C. Nascimento, J. S. Marques, and M. A. T. Figueiredo, "Discriminative model selection using a modified Bayesian criterion: Application to trajectory modeling," in *Proc. ICIP*, Sep. 2011, pp. 1429–1432.

[29] T. E. Boult, R. J. Micheals, X. Gao, and M. Eckmann, "Into the woods: Visual surveillance of noncooperative and camouflaged targets in complex outdoor settings," in *Proc. IEEE*, vol. 89, no. 10, pp. 1382–1402, Oct. 2001.



Jacinto C. Nascimento (S'00–M'06) received the Electrical Engineering degree from the Instituto Superior de Engenharia de Lisboa, Lisboa, Portugal, in 1995, and the M.Sc. and Ph.D. degrees from the Instituto Superior Técnico (IST), Technical University of Lisbon, Lisbon, in 1998 and 2003, respectively. He is currently an Assistant Professor with the Department of Informatics and Computer Engineering, IST, and a Researcher with the Institute for Systems and Robotics. He has authored 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, and general visual object classification.

Miguel Barão received the Electrical Engineering, M.Sc., and Ph.D. degrees from the Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal, in 1996, 2000, and 2008, respectively. He is currently an Assistant Professor with the Department of Informatics, Universidade de Évora, Évora, Portugal, and a Researcher with the Control of Dynamical Systems Group, Instituto de Engenharia de Sistemas e Computadores-Investigação e Desenvolvimento, Lisbon. 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, and probabilistic geometric control. He has authored or co-author several journal and conference papers in control.

Jorge S. Marques received the Electrical Engineering, Ph.D., and the Aggregation degrees from the Technical University of Lisbon, Lisbon, Portugal, in 1981, 1990, and 2002, respectively. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Instituto Superior Técnico, Technical University of Lisbon, and a Researcher with the Institute for Systems and Robotics. He has authored over 150 papers in international journals and conferences, and is the author of the book entitled *Pattern Recognition: Statistical and Neural Methods–2nd Edition* (IST Press, 2005, in Portuguese). He was the Co-Chairman of the International Association for Pattern Recognition Iberian Conference on Pattern Recognition and Image Analysis (2005), the President of the Portuguese Association for Pattern Recognition (2001–2003), and an Associate Editor of the *Statistics and Computing* (Springer) journal. His research interests are in the areas of statistical image processing, shape analysis, and pattern recognition.

João M. Lemos received the Ph.D. degree from the Instituto Superior Técnico (IST), Technical University of Lisbon, Lisbon, Portugal, in 1989, after extensive period of research work with the University of Florence, Florence, Italy, and a period of experimental work with the Department of Chemical Engineering, Imperial College of Science, Technology and Medicine, London, U.K., and became an Associate Professor in 1997. He is currently a Full Professor (Professor Catedrático) of Systems Decision and Control with IST and a Researcher with the Instituto de Engenharia de Sistemas e Computadores-Investigação e Desenvolvimento (INESC-ID), Lisbon, where he has led the Control of Dynamic Systems Group since 1990. He has been the Chairman of the Department of Electrical and Computer Engineering with IST and a Coordinator of the Post-Graduation Program in Electrical and Computer Engineering, and the Chairman of the Scientific Board of INESC-ID since 2007. He has 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, modeling 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, and biochip temperature control), automotive systems, and water delivery canals.