A SPARSE APPROACH TO PEDESTRIAN TRAJECTORY MODELING USING MULTIPLE MOTION FIELDS

Catarina Barata *Jacinto C. Nascimento* *Jorge S. Marques*

Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, Portugal

ABSTRACT

This paper proposes a novel methodology to describe the trajectories performed by pedestrians in long-range surveillance scenarios. The proposed approach describes the trajectories by using sparse motion/vector fields together with a space-varying switching mechanism embedded in a Hidden Markov Model framework. Despite the diversity of motion patterns that may occur in a given scenario, the observed trajectories do not lie in the entire surveilled area. Instead, they are constrained to patterns corresponding to typical motions. To achieve a compact representation, we propose a sparse model estimated using the ℓ_1 norm applied to the log prior distribution of the vector fields. Experimental evaluation is conducted in real scenarios, and testify the usefulness of the proposed approach in modeling typical trajectories that occur in a far-field surveillance setup.

***Index Terms*—** Motion estimation, pedestrian motion analysis, motion fields, sparse representation.

1. INTRODUCTION

Describing the trajectories performed by pedestrians constitutes a valuable step to be incorporated in a surveillance setup, since such information helps to understand what people are doing in the scene (*i.e.*, recognize human activities and interactions [1,2]), and what actions should be conducted from those observations. The topic of interpreting and classifying human activities has plateaued in the last few years, given the vast number of applications, *e.g.*, intelligent environments [3,4], human machine interaction [5], sports analysis [6], and surveillance [7–9]. It has been recently demonstrated the usefulness of vector fields, not only in surveillance settings [7–9], but also in other problems *e.g.,* analysis of hurricane data, GPS tracks of people and vehicles, and cellular radio handoffs [10]. The focus of this paper is to model the trajectories using a set of sparse motion/vector fields.

Vector fields were used previously [7], where each trajectory was represented by a sequence of segments, each of which generated by one vector field. Switching between models occur at any point in the image domain with a probability that depends on the spatial location. This model is flexible enough to represent a wide variety of motion patterns. The expectation-maximization (EM) algorithm was applied to learn the model parameters from the observed trajectories. However, the drawback with the above approach is that we obtain a "dense" representation of the vector fields in the image domain. By dense, we mean that we are estimating the vector fields in regions where no observations were acquired. In this paper we are able to circumvent this limitation by estimating the motion/vector fields only on regions where observations exist. This is accomplished by enforcing a sparse solution using an appropriate prior as it is discussed in the paper.

This work was supported by the FCT project and plurianual funding: [PTDC/EEIPRO/0426/2014], [UID/EEA/50009/2013].

Fig. 1. Pedestrian trajectories at IST campus, Lisbon

The rest of the paper is organized as follows. Section 2 discusses related work. Sections 3 and 4 describe the trajectory model and the motion field model. Section 5 address the motion field estimation using a MAP criterion and section 6 discusses the prior. Section 7 discusses the optimization procedure. Section 8 presents experimental results and section 9 draws the main conclusions.

2. RELATED WORK

Image processing algorithms have immensely progressed in the past decades. This was achieved using *better image models*. One of the research directions that has undoubtedly contributed for this progress, is the use of the sparse representation techniques. Sparse representation techniques allow to obtain robust and reliable models that have been sucessfully used in several problems, ranging from denoising, restauration, to sampling and inverse problems for reconstruction [11]. Related work in sparse methodologies has been proposed in finding low complexity algorithms that yield solutions to the aforementioned problems. Two classical families of algorithms are the *matching pursuit* (MP) [12,14] and the FOCal Underdetermined System Solver (FOCUSS). The MP algorithms are based on suboptimal forward sequential algorithms [12–14], that work by successively adding vectors until the representation of the desired signal is achieved. Other techniques are based on optimization methodologies that maximize sparsity, such as the ℓ_1 norm [15], or the more general $\ell_{(\rho<1)}$ explored by *FOCal Underdetermined System Solver* (FOCUSS) [16]. As it will be described herein, sparse methodologies have enlarged its domains of application.

In the context of the present paper, two main concepts have been adopted. In one hand, the use of trajectories has been recently proposed for activity monitoring. For instance, in [9] a parametric Dirichlet process is applied for trajectory clustering, and where online classification is achieved using a particle based filtering approach for abnormality detection. Dense trajectories for human mo-

tion recognition has also been proposed [17], where the main idea is to perform a dense sampling on feature points in each frame, and then perform tracking based on optical flow. Sparse based methods, on the other hand, have played an important role in human activity recognition. In [18] is proposed an algorithm for learning sparse, spatiotemporal features applying the resulting sparse codes to the problem of activity recognition.

In this paper, we propose to join the aforementioned concepts by proposing a new *sparse* approach to model the *trajectories* performed by pedestrians. To our knowledge, the use of these two concepts was never used to estimate the vector fields for video surveillance purposes. More specifically, we consider three prior models that introduce different sparsity levels. It will be shown that the proposed approach is useful for modelling of trajectories in a surveillance setup.

3. TRAJECTORY MODEL

Let us consider a pedestrian trajectory in the image plane, $x = (x_1, \ldots, x_L)$, where $x_t \in [0,1]^2$ denotes the pedestrian position at discrete time t (see Fig. 1). We will assume that the pedestrian motion is driven by K motion fields $T_k : [0,1]^2 \rightarrow \mathbb{R}^2$ as suggested in [7]. It is assumed that each trajectory is generated by a bank of switched dynamical models,

$$x_t = x_{t-1} + T_{k_t}(x_{t-1}) + w_t, \tag{1}$$

where $k_t \in \{1, \ldots, K\}$ denotes the active motion field at time t and $w_t \sim N(0, \sigma^2 I)$ is a white random perturbation.

We will also assume that switching between active fields depends on the pedestrian position and is modeled as a first order Markov process. Therefore,

$$P(k_t = j | k_{t-1} = i, x_{t-1}) = B_{ij}(x_{t-1}), \tag{2}$$

where $B(x) = \{B_{ij}(x), i, j \in \{1, \ldots, K\}\}$ is a space-varying matrix of switching probabilities that depends on the pedestrian position. Therefore, $B(x)$ is a stochastic matrix for each admissible position, $x \in [0,1]^2$.

If the sequence of switching labels $k = (k_1, \ldots, k_L)$ is known, the joint probability $p(x, k)$ can be easily obtained

$$p(x,k) = p(x_1, k_1) \prod_{t=2}^{L} p(x_t, k_t | x_{t-1}, k_{t-1}) \tag{3}$$

$$= p(x_1, k_1) \prod_{t=2}^{L} p(x_t | k_t, x_{t-1}) p(k_t | k_{t-1}, x_{t-1}), \tag{4}$$

and taking log we obtain,

$$\log p(x,k) = \log p(x_1, k_1) - (L-1)\log(2\pi\sigma^2) - \\ - \frac{1}{2\sigma^2} \sum_{t=2}^{L} \|x_t - x_{t-1} - T_{k_t}(x_{t-1})\|_2^2 + \\ + \sum_{t=2}^{L} \log B_{k_{t-1}, k_t}(x_{t-1}). \tag{5}$$

In this paper we will assume that $B_{i,j}(x)$ is known and we will focus on the estimation of the motion fields T_k from the training data. The estimation of $B(x)$ is addressed elsewhere [19]. But first, we need to discuss how are the motion fields represented.

4. MOTION FIELD REPRESENTATION

Motion fields can be represented in several ways *e.g.*, using parametric models [8] or Gaussian processes [9]. In this paper, we will assume that motion fields are freely specified at the nodes of a regular grid on the image plane, $\mathcal{G} = \{g_i \in [0,1]^2, i = 1, \ldots, N\}$, and interpolated in other image points $x \notin \mathcal{G}$. Let $T_k \in \mathbb{R}^{N \times 2}$ be a matrix with the values of k-th motion field at the grid nodes. For $x \notin \mathcal{G}$, the motion field is obtained by bilinear interpolation as follows

$$T_k(x) = \Phi(x) T_k, \tag{6}$$

where $\Phi(x)$ is a $N \times 1$ sparse matrix of interpolation coefficients (only 4 coefficients will be non-zero).

The parameters to be estimated are, thus, the velocities of the K motion fields at the grid nodes: $T = (T_1, \ldots, T_K)$.

5. MAP ESTIMATION

We will now assume that we have observed a training set of S trajectories, $\mathcal{X} = \{x^{(1)}, \ldots, x^{(S)}\}$, and wish to estimate the motion fields. The MAP estimate of the motion fields parameters is given by

$$\hat{T} = \arg\max_{T} \left[\log p(\mathcal{X}|T) + \log p(T)\right], \tag{7}$$

where $\log p(T)$ is the log prior distribution. Assuming that different motion fields have independent priors, we obtain $\log p(T) = \sum_{k=1}^{K} \log p(T_k)$.

Unfortunately, the likelihood function $p(\mathcal{X}|T)$ cannot be analytically computed. The complete likelihood function $p(\mathcal{X}, \mathcal{K}|T)$ can be easily obtained from (5), but the marginalization

$$p(\mathcal{X}|T) = \sum_{\mathcal{K}} p(\mathcal{X}, \mathcal{K}|T), \tag{8}$$

involves a sum over all the admissible label sequences, $\mathcal{K}$, and it is not feasible.

To overcome this difficulty we will adopt the expectation-maximization (EM) method using the auxiliary function

$$U(T, T') = E\left\{\log p(\mathcal{X}, \mathcal{K}|T) | \mathcal{X}, T'\right\} + \log p(T), \tag{9}$$

where T' denotes the most recent estimate of T. This function can be analytically computed since the marginalization was replaced by an expectation.

Using (5) and neglecting constant terms, we obtain

$$U(T, T') = \sum_{k=1}^{K} \log p(T_k) - \\ - \frac{1}{2\sigma^2} \sum_{s=1}^{S} \sum_{t=2}^{L_s} \sum_{k=1}^{K} w_k^{(s)}(t) \|x_t^{(s)} - x_{t-1}^{(s)} - T_k(x_{t-1}^{(s)})\|_2^2, \tag{10}$$

where $w_i^{(s)}(t) = P(k_t^{(s)} = i | x^{(s)}, T')$ is the probability of assigning the $t-th$ observation to model i, in sequence s. This is computed in the E-step.

The M-step of the EM method involves the maximization of (10) that can be split into K independent subproblems *i.e.*, one per motion field by maximizing

$$U_k(T, T') = -\frac{1}{2\sigma^2} \sum_{s=1}^{S} \sum_{t=2}^{L_s} w_k^{(s)}(t) \|x_t^{(s)} - x_{t-1}^{(s)} - T_k(x_{t-1}^{(s)})\|_2^2 \\ + \log p(T_k), \tag{11}$$

and each of them can be separately computed for x and y components of the motion fields, provided that $\log p(T_k)$ can be split into the sum of two functions for x and y velocity components.

6. PRIOR DISTRIBUTION

The prior distribution, $p(T)$, is used to incorporate *a priori* knowledge about the parameters, keeping the estimation process simple. We will assume that each motion field has two properties:

- smoothness: for each pair of neighboring grid nodes $x_{g1}, x_{g2} \in \mathcal{G}$, the velocity difference $T_k(x_{g1}) - T_k(x_{g2})$ should be small;
- small coefficients: in most grid nodes $x_g \in \mathcal{G}$, the velocity $T_k(x_g)$ should be small.

The second condition may lead to sparse solutions if the l_0 or l_1 norms are used. To enforce these conditions we will consider the following logarithm of the prior (normalization constants were discarded)

$$\log p(T_k) = \alpha \|\Delta T_k\|_2^2 + \beta \|T_k\|_p^p, \tag{12}$$

where Δ is an operator that computes all differences between velocities of neighboring nodes, $p \in \{1, 2\}$, and $\|.\|_p$ denotes the pth norm of a matrix. We will consider three priors:

1. $\beta = 0$ (the choice adopted in [7]);
2. $\beta \neq 0, p = 2$ squared error penalty of the coefficients and
3. $\beta \neq 0, p = 1$, l_1 norm enforcing sparseness in the set of coefficients.

7. OPTIMIZATION

The cost functional (10) can be split into K independent cost functionals (one per motion field) and each of these can be split again into two simpler cost functionals: one for vertical velocity field and one for horizontal velocity field. Each of these $2K$ optimization problems can be written in the form

$$\min_{u_k} \|W_k(v - \phi u_k)\|_2^2 + \alpha \|\Delta u_k\|_2^2 + \beta \|u_k\|_p^p, \tag{13}$$

where v includes the pedestrian velocities in all trajectories, u_k contains the grid coefficients (horizontal or vertical velocities) for each field $k = 1, 2, ..., K$, $p \in \{1, 2\}$, and W_k is a diagonal weight matrix of the form

$$W_k = \begin{bmatrix} \sqrt{w_k^{(1)}(2)}I & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \dots & \vdots & \dots & \vdots & \dots & \vdots \\ 0 & \dots & \sqrt{w_k^{(1)}(L_s)}I & \dots & 0 & \dots & 0 \\ \vdots & \dots & \vdots & \dots & \vdots & \dots & \vdots \\ 0 & \dots & 0 & \dots & \sqrt{w_k^{(S)}(2)}I & \dots & 0 \\ \vdots & \dots & \vdots & \dots & \vdots & \dots & \vdots \\ 0 & \dots & 0 & \dots & 0 & \dots & \sqrt{w_k^{(S)}(L_s)}I \dots \end{bmatrix}.$$

This is a convex cost function that is solved by using CVX package for convex programs [20]

8. EXPERIMENTAL RESULTS

The algorithm was applied to model the motion of pedestrians in the Campus data set [7]. This data set contains the trajectories of 134 pedestrians (14308 points) acquired by a static camera at IST Lisbon. Fig. 2 (top) shows the trajectories extracted superimposed on one of the Campus images. The images were first transformed by an homography, in order to compensate the distortion caused by the perspective projection. Then, the pedestrians were detected and tracked (see Fig. 2 (bottom left)). These are the trajectories that were used to estimate the motion fields. We have transformed the trajectories back by the inverse homography for display purposes only (see Fig. 2 (top)).

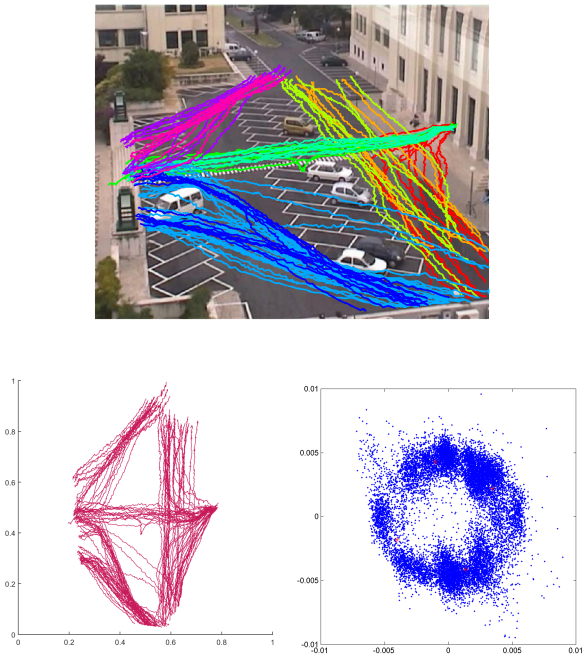

Fig. 2. Trajectories extracted from the IST Campus sequence, each color denotes a different class-trajectory (top). Trajectories compensated by an homography (bottom left) and pedestrian velocities and cluster centers obtained with a Gaussian mixture model (bottom right).

We set the number of fields to $K = 4$ (roughly North, South, East and West) and used a grid of 21×21 nodes. The learning algorithm was initialized using uniform motion fields, with the field values obtained by clustering the pedestrian velocities using Gaussian mixture models (see Fig. 2 (bottom right)). This initialization was considered to be robust since the algorithm always converged to acceptable field configurations in all the trials.

We tested three priors: i) $\beta = 0$; ii) $\beta = 0.2, p = 2$, iii) $\beta = 0.02, p = 1$, and $\alpha = 0.2$ in all experiments (α and β were empirically chosen). Typical results obtained with these priors are shown in Figs. 3-5. They all approximate the pedestrian velocity well. This can be measured by computing the energy of the prediction error as defined in (10). However, the first prior ($\beta = 0$) extends the motion field to the whole image, including regions where no data is observed. In such regions the model makes no sense since no information was observed there. In addition, the estimated motion fields are difficult to interpret in this case. When we penalize the coefficients with a squared Euclidean norm, the previous problem is reduced but the estimated fields still tend to occupy most of the image with non-zero velocity estimates. The third prior is the one that performs best. It provides non-zero estimates in the regions where data is concentrated and zero velocity estimates in the regions where there is no data.

To characterize the estimated fields, we computed the amount of overlap between all the fields as follows

$$\mathcal{O} = \frac{2}{K(K-1)} \sum_{i=1}^{K} \sum_{j=1}^{i-1} \mathcal{O}_{ij} \tag{14}$$

where $\mathcal{O}_{ij}$ measures the overlap between fields i and j, *i.e.*, the per-

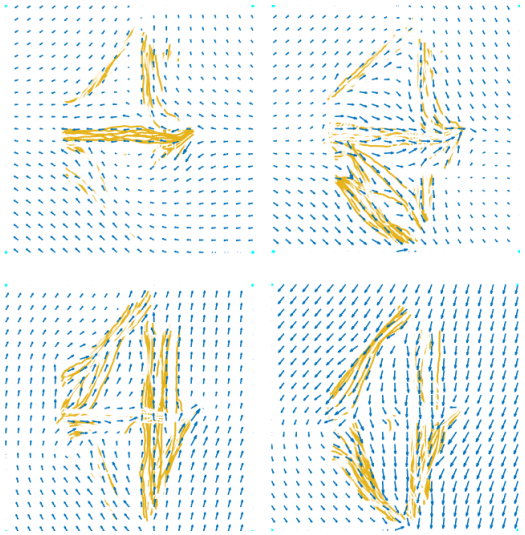

Fig. 3. Learned motion fields with prior 1 ($\beta = 0$) and trajectory segments associated to each field

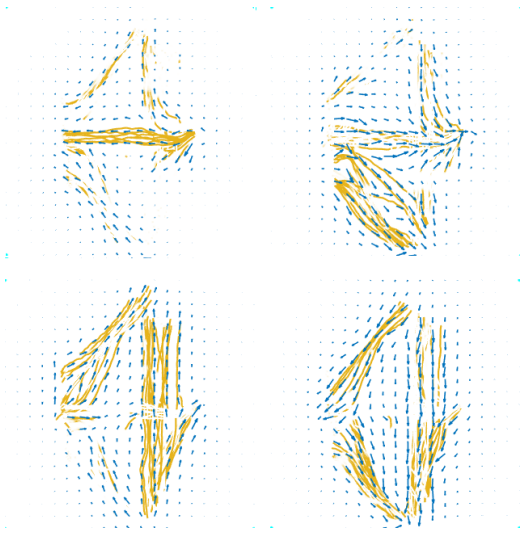

Fig. 4. Learned motion fields with prior 2 ($\beta = 0.2, p = 2$) and trajectory segments associated to each field

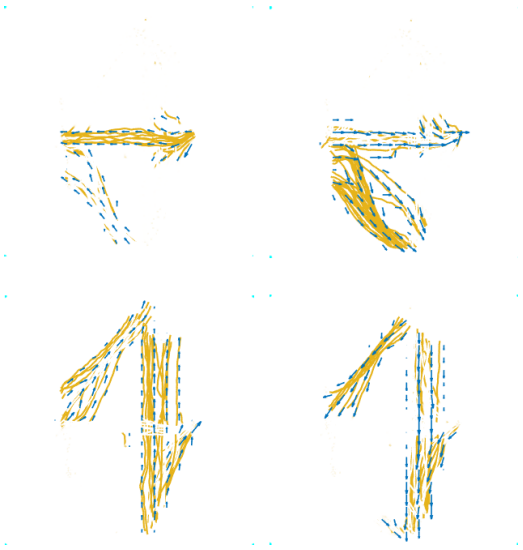

Fig. 5. Learned motion fields with prior 3 ($\beta = 0.02, p = 1$) and trajectory segments associated to each field

Table 1. Percentage of field overlap associated to each prior

	Prior 1	Prior 2	Prior 3
Overlap	99.8%	44.7%	16.6 %

centage of grid nodes in which both fields have values greater then a small threshold. Table 1 shows the experimental results. The first prior leads to velocity field estimates that cover the whole image, showing a full overlap (100%) even in regions where there are no data. The field estimates in such regions is meaningless. On the contrary, the third prior (l_1 - norm) leads to sparse vector fields that have a much smaller overlap (17%), since the field estimates are zero if there is no data in a region. The interpretation of the velocity field estimates is also much easier. The second prior (squared $l2$ norm) lies in the middle. The best results are obtained with the l_1 regularization.

9. CONCLUSIONS

This paper studies the effect of the prior on the motion fields estimates. It is shown that a prior based on the ℓ_1 norm enforces sparseness on the multiple motion field estimates and has significant advantages. First, the estimated fields are zero in regions where there is no data to support them. This is important because it does not make sense to use motion fields in regions where there is no training data. Second, it also simplifies the motion fields and make them more easy to interpret. Third, motion fields overlap much less. The proposed algorithm is robust and leads to meaningful description of the data.

10. REFERENCES

[1] J. Aggarwal and S. Park, "Human motion: Modeling and recognition of actions and interactions," in Proc. 3-D Data Process., Vis. Trans. Conf., Sep. 2004, pp. 640-647

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

[3] D. Hu, S. Pan, V. Zheng, N. Liu, and Q. Yang, "Real world

activity recognition with multiple goals," in Proc. ACM 10th Int. Conf. Series, 2008, pp. 30-39.

[4] "Intelligent Environments, Methods, Algorithms and Applications", D. Monekosso, P. Remagnino and Y. Kuno Editors, Springer, 2008.

[5] M. Pantic, A. Pentland, A. Nijholt, and T. Huang, "Human computing and machine understanding of human behavior: A survey," in Proc. 8th Int. Artif. Intell. Human Comput. Conf., 2007, pp. 239-248.

[6] M. Perse, M. Kristan, S. Kovacic, G. Vuckovic, and J. Persa, "A trajectory-based analysis of coordinated team activity in a basketball game," Comput. Vis. Image Understand., vol. 113, no. 5, pp. 612-621, 2009.

[7] J. C. Nascimento, M. A. T. Figueiredo, J. S. Marques. "Activity recognition using mixture of vector fields", *IEEE Trans. on Image Processing*, vol. 22, no. 5, pp. 1712 - 1725, 2013.

[8] J. C. Nascimento, J. S. Marques and J. M. Lemos, "Modeling and classifying human activities from trajectories using a class of space-varying parametric motion fields", *IEEE Trans. on Image Processing*, vol. 22, no. 5, pp. 2066-2080, 2013.

[9] V. Bastani, L. Marcenaro and C. S. Regazzoni, "Online Nonparametric Bayesian Activity Mining and Analysis From Surveillance Video", *IEEE Trans. on Imag. Proc.*, vol. 25, pp. 2089-2102, 2016.

[10] N. Ferreira, J. T. Klosowski, C. E. Scheidegger and C. T. Silva, "Vector Field k-Means: Clustering Trajectories by Fitting Multiple Vector Fields", Eurographics Conference on Visualization, vol. 32 no. 3, pp. 201-210, 2013.

[11] M. Elad, M. A.T. Figueiredo and Yi Ma, "On the Role of Sparse and Redundant Representations in Image Processing", Proc. of the IEEE, special issue on applications of sparse representation and compressive sensing, 2010.

[12] B. K. Natarajan, "Sparse approximate solutions to linear systems," SIAM J. Comput., vol. 24, no. 2, pp. 227-234, 1995

[13] E. S. Cheng, S. Chen, and B. Mulgrew, "Efficient computational schemes for the orthogonal least squares learning algorithm," IEEE Trans. Signal Process., vol. 43, no. 1, pp. 373-376, Jan. 1995

[14] S. F. Cotter, J. Adler, B. D. Rao, and K. Kreutz-Delgado, "Forward sequential algorithms for best basis selection," Proc. Inst. Elect. Eng. Vision, Image, Signal Process., vol. 146, no. 5, pp. 235-244, 1999.

[15] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," IEEE Trans. Signal Process., vol. 47, no. 1, pp. 187-200, 1999.

[16] K. Kreutz-Delgado and B. D. Rao "Convex/schur-convex (CSC) log-priors and sparse coding," in Proc. 6th Joint Symp. Neural Comput., 1999.

[17] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition", IJCV, 103(1):60-79, 2013.

[18] T. Dean, G. Corrado and R. Washington, "Sparse Spatiotemporal Coding for Activity Recognition", Tech. Report *http://www.cs.brown.edu/research/pubs/techreports/reports/CS-10-02.html*, Brown University, 2010.

[19] J. C. Nascimento, M. Barão, J. S. Marques and J. M. Lemos, "Information Geometric Algorithm for Estimating Switching Probabilities in Space-Varying HMM", IEEE Trans. Image Process., Vol. 23, no 12, 5263-5273, 2014.

[20] M. Grant and S. Boyd, CVX: Matlab Software for Disciplined Convex Programming, version 2.1, http://cvxr.com/cvx, 2014