Discovering motion hierarchies via tree-structured coding of trajectories

Juan-Manuel Pérez-Rúa[1]
juanmanuel.perezrua@technicolor.com
Tomas Crivelli[1]
http://www.technicolor.com/en/tomas-crivelli
Patrick Pérez[1]
http://www.technicolor.com/en/patrick-perez
Patrick Bouthemy[2]
patrick.bouthemy@inria.fr

[1] Technicolor R&I
Cesson Sévigné, France
[2] Inria
Centre Rennes
Bretagne Atlantique, France

Figure 1: Hierarchical organization of visual motions in a natural scene.

The dynamic content of physical scenes is largely compositional, that is, the movements of the objects and of their parts are hierarchically organized and relate through composition along this hierarchy. This structure also prevails in the apparent 2D motion that a video captures (see Fig.1). Accessing this visual motion hierarchy is important to get a better understanding of dynamic scenes and is useful for video manipulation.

Early works in biological vision found that visual systems decompose objects into parts through the analysis of motion nesting [3]. Johansson showed in particular that removing the motion of the main body from the image reveals the distinctive motion of its parts. Along the same lines, Gershman *et al*. [1] have recently proposed a computational model that can decompose dynamic sensory data into a hierarchy of components. The hierarchical decomposition of visual motion information is thus clearly identified as a key step in complex biological vision systems.

In this paper, we propose to investigate these ideas in the context of video analysis for the benefit of complex tasks like video understanding and parsing. We aim at capturing the hierarchical video representation through learned, tree-structured sparse coding of point trajectories. We leverage this new representation within an unsupervised clustering scheme to partition hierarchically the trajectories into meaningful groups.

Given an input video sequence of $M+1$ frames and N input point trajectories extracted from it $(\mathbf{x}^n_{0:M}) \in \mathbb{R}^{2\times(M+1)}$, $n = 1\cdots N$, we define the data matrix $X \in \mathbb{R}^{2M\times N}$ as:

$$X = \begin{bmatrix} \Delta\mathbf{x}^1_1 & \Delta\mathbf{x}^2_1 & \cdots & \Delta\mathbf{x}^N_1 \\ \Delta\mathbf{x}^1_2 & \Delta\mathbf{x}^2_2 & \cdots & \Delta\mathbf{x}^N_2 \\ \vdots & \vdots & \cdots & \vdots \\ \Delta\mathbf{x}^1_M & \Delta\mathbf{x}^2_M & \cdots & \Delta\mathbf{x}^N_M \end{bmatrix}, \quad (1)$$

where $\Delta\mathbf{x}^n_m = \mathbf{x}^n_m - \mathbf{x}^n_{m-1}$. In this matrix, each column stems for the sequence of displacements along one trajectory.

A powerful way to discover multiple structures in such data is through sparse coding with a learned dictionary. One seeks an approximate decomposition $X \approx DA$ into a *dictionary* matrix $D = [\mathbf{d}_1 \cdots \mathbf{d}_K] \in \mathbb{R}^{2M\times K}$, and a *sparse representation* $A = [\alpha_1 \cdots \alpha_N] \in \mathbb{R}^{K\times N}$ of the input data. Such a sparse decomposition can be achieved by solving the optimization problem:

$$\arg\min_{D,A} \|X - DA\|_2^2 \quad \text{sb.t. } \|\alpha_n\|_0 \le s,\ \forall n \quad \text{and} \quad \|\mathbf{d}_k\|_2 = 1,\ \forall k \quad (2)$$

The previous formulation, however, does not enforce any structure among the atoms of the dictionary and on the associated codes. We then re-formulate the problem so that the dictionary and the encoding are constrained in certain way by a tree structure. In particular, we want the movement of a given scene element to be represented *only with dictionary atoms stemming from a same branch of the tree*. For a given rooted tree $\mathcal{T}$ of K nodes numbered in level-order we want to learn a dictionary $D = [\mathbf{d}_{1:K}] \in \mathbb{R}^{2M\times K}$ of K trajectory atoms organized according to this tree structure, together with the corresponding matrix $A = [\alpha_{1:N}] \in \mathbb{R}^{K\times N}$ of sparse codes. To this end, we consider the following constrained minimization problem:

$$\arg\min_{D,A} \|X - DA\|_2^2, \quad \text{sb.t. } \alpha_n \in \mathcal{A}(\mathcal{T}),\ \forall n \quad \text{and} \quad \|\mathbf{d}_k\|_2 = 1,\ \forall k, \quad (3)$$

where $\mathcal{A}(\mathcal{T}) \subset \mathbb{R}^K$ is the set of tree-structured codes defined as:

$$\mathcal{A}(\mathcal{T}) = \{\alpha \in \mathbb{R}^K : \mathrm{supp}(\alpha) = \mathrm{anc}(\mathrm{k}(\alpha))\}, \quad (4)$$

where $\mathrm{anc}(k)$ denotes the ancestor set of node k in $\mathcal{T}$ (the nodes, including itself, that form the unique path from k to root node 1), $\mathrm{supp}(\alpha)$ is the support of α, that is the index set of its non-zero entries, and $\mathrm{k}(\alpha) = \max(\mathrm{supp}(\alpha))$ stands for the last atom in the code.

We use the K-SVD algorithm to solve the dictionary learning problem (3), but we modify the orthogonal matching pursuit (OMP) encoding part in order to respect the constraint in (4). The algorithm is shown in detail in the paper.

Having all trajectories encoded in A according to the learned tree-structured dictionary D already provides a hierarchical partitioning of the trajectories one by gathering, for each node k in the tree, all trajectories such that $k \in \mathrm{anc}(\mathrm{k}(\alpha))$. Unfortunately, such partitions are noisy since the above dictionary learning and sparse coding are not explicitly geared toward a clustering task.

In the same spirit as spectral clustering that conducts final K-means clustering on spectrally encoded data vectors, we can cluster the trajectories based on their codes α_ns, with hierarchical K-means in our case. This already provides cleaner partitions. Drawing inspiration from Jiang *et al*. [2] who combine dictionary learning with supervised learning of linear classifiers over codes, we can go one step further: given a current hierarchical clustering of trajectories' codes, we can update our tree-structured dictionary and iterate. As will appear in the experiments, this procedure further improves the quality of track clusters. At each iteration, the dictionary learning problem to solve becomes

$$\arg\min_{D,Y,A} \|X - DA\|_2^2 + \lambda\|Q - YA\|_2^2, \quad \text{sb.t. } \alpha_n \in \mathcal{A}(\mathcal{T}),\ \forall n \quad (5)$$

where $Q \in \{0,1\}^{K\times N}$ is the binary matrix associated to the current hierarchical clustering of tracks (each of its columns belongs to $\mathcal{A}(\mathcal{T})$) and λ is a positive parameter that controls the balance between reconstruction and clustering terms. This new objective can also be optimized with K-SVD, as shown in [2]. We show in the paper how to use our algorithm for hierarchical and flat clustering of trajectories taken from video sequences. We also analyze hierarchical motion patterns that are common in human activities like walking and jumping. We show that our algorithm is capable, up to some level, to separate nested and independent motions.

[1] Samuel J Gershman, Joshua B Tenenbaum, and Frank Jäkel. Discovering hierarchical motion structure. *Vision Research*, 2015. In press.

[2] Zhuolin Jiang, Zhe Lin, and Larry S Davis. Learning a discriminative dictionary for sparse coding via label consistent K-SVD. In *CVPR, Colorado Springs*, 2011.

[3] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Attention, Perception, & Psychophysics*, 14(2):201–211, 1973.

[4] Peter Ochs, Jagannath Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1187–1200, 2014.