Kernelized Covariance for Action Recognition

Jacopo Cavazza[1,2], Andrea Zunino[1,2], Marco San Biagio[1], Vittorio Murino[1,3]

[1] Pattern Analysis & Computer Vision – Istituto Italiano di Tecnologia, Via Morego 30, 16163, Genova, Italy

[2] Università degli Studi di Genova – Dipartimento di Ingegneria Navale, Elettrica, Elettronica e delle Telecomunicazioni, Via all'Opera Pia, 11A, 16145, Genova, Italy

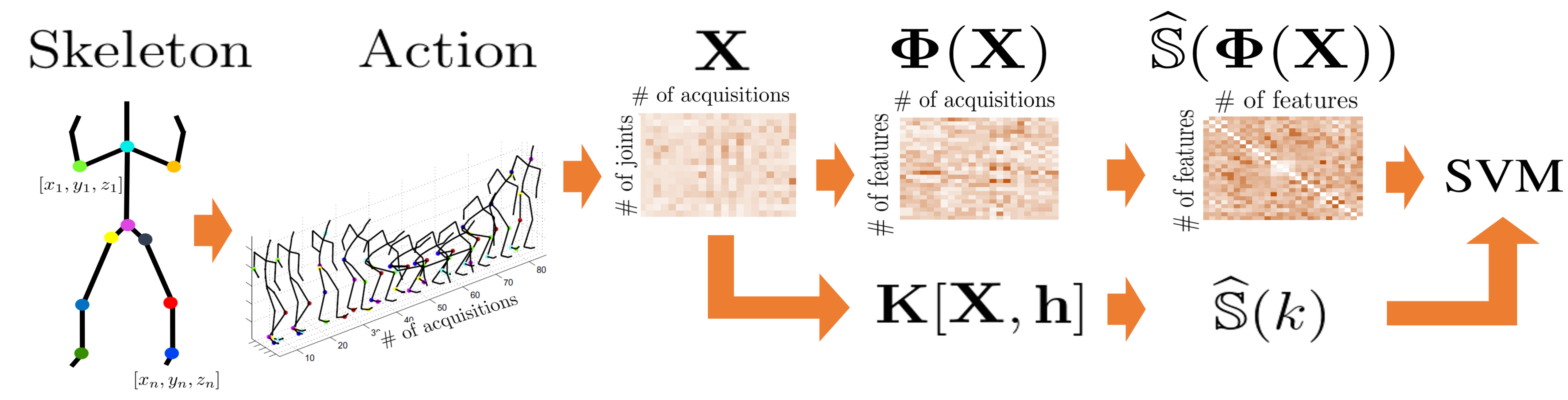
[3] Università di Verona – Dipartimento di Informatica, Strada le Grazie 15, 37134, Verona, Italy

{jacopo.cavazza,andrea.zunino,marco.sanbiagio,vittorio.murino}@iit.it

Abstract

- Originally devised as an image descriptor [5], the covariance matrix is powerful in correlating skeletal joints across time for action recognition [2, 10].
- As the main limitation, covariance can only capture *linear* mutual relationships.
- In this work, we extend **covariance to model arbitrary, non-linear relationships** by recovering the applicability of the **kernel trick** and consequently **avoiding any preliminary feature encoding** of the raw data.

Pipeline



Recovering the Kernel Trick for the Covariance Representation

Each action instance is represented through $\mathbf{X} = [\mathbf{x}(1), \dots, \mathbf{x}(T)] \in \mathbb{R}^{3n \times T}$, collecting the n joints positions $\mathbf{x}_1(t), \mathbf{x}_2(t) \dots, \mathbf{x}_n(t)$ at each timestamp $t = 1, \dots, T$, being $\mathbf{x}_i(t) = [x_i(t) y_i(t) z_i(t)]$. We introduce the sampling covariance operator, defined as

$$\widehat{\mathbb{S}}(\mathbf{X}) = \frac{1}{T-1} \sum_{t=1}^{T} (\mathbf{x}(t) - \boldsymbol{\mu})(\mathbf{x}(t) - \boldsymbol{\mu})^\top, \text{ where } \boldsymbol{\mu} = \frac{1}{T} \sum_{s=1}^{T} \mathbf{x}(s), \text{ which rewrites } \widehat{\mathbb{S}}(\mathbf{X}) = \mathbf{X}\mathbf{P}\mathbf{X}^\top \text{ once defined } \boxed{\mathbf{P} = \frac{1}{T-1}\left(\frac{1}{T}\mathbf{I} - \mathbf{1}\right)}$$

As to model non-linear correlations within the data through covariance, $\widehat{\mathbb{S}}(\Phi(\mathbf{X})) = \mathbf{\Phi}(\mathbf{X})\mathbf{P}\mathbf{\Phi}(\mathbf{X})^\top$ is computed in terms of an explicit feature map $\Phi \colon \mathbb{R}^{3n} \to \mathcal{H}$ which has to be applied to the whole data matrix $\mathbf{X}$, obtaining $\mathbf{\Phi}(\mathbf{X}) = [\Phi(\mathbf{x}(1)), \dots, \Phi(\mathbf{x}(T))] \in \mathcal{H} \times \dots \times \mathcal{H}$.
We posit that $\widehat{\mathbb{S}}(\Phi(\mathbf{X}))$ can be computed in terms of the kernel $k(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle_{\mathcal{H}}$ *only*, if provided the existence of $\mathbf{h}_i$ such that $\Phi(\mathbf{h}_i) = \mathbf{e}_i$ for any i. Indeed,

$$\widehat{\mathbb{S}}_{ij}(\mathbf{\Phi}(\mathbf{X})) = \sum_{s=1}^{T}\sum_{t=1}^{T} \Phi_i(\mathbf{x}(s))\mathbf{P}_{st}\Phi_j(\mathbf{x}(t)) = \sum_{s=1}^{T}\sum_{t=1}^{T} \langle \Phi(\mathbf{x}(s)), \mathbf{e}_i \rangle_{\mathcal{H}} \mathbf{P}_{st} \langle \Phi(\mathbf{x}(t)), \mathbf{e}_j \rangle_{\mathcal{H}} = \sum_{s=1}^{T}\sum_{t=1}^{T} \underbrace{\langle \Phi(\mathbf{x}(s)), \Phi(\mathbf{h}_i) \rangle_{\mathcal{H}}}_{k(\mathbf{x}(s), \mathbf{h}_i)} \mathbf{P}_{st} \underbrace{\langle \Phi(\mathbf{x}(t)), \Phi(\mathbf{h}_j) \rangle_{\mathcal{H}}}_{k(\mathbf{x}(t), \mathbf{h}_j)} = \widehat{\mathbb{S}}_{ij}(k).$$

As the main theoretical contribution of our work, we show that the assumption $\Phi(\mathbf{h}_i) = \mathbf{e}_i$ can be fulfilled by a particular class of random feature maps related to a Taylor kernel function $k(\mathbf{x}, \mathbf{z}) = \sum_{\ell=0}^{\infty} a_\ell \langle \mathbf{x}, \mathbf{z} \rangle^\ell$, being $a_\ell > 0$ for any ℓ. For concreteness, let us fix $k(\mathbf{x}, \mathbf{z}) = \exp(\gamma \cdot \langle \mathbf{x}, \mathbf{y} \rangle)$ for a given $\gamma > 0$.

Random Approximated Feature Map [3]

$\mathbf{\Psi} \colon \mathbb{R}^{3n} \to \mathcal{H}, \; \Psi_i(\mathbf{x}) \overset{i.i.d}{\sim} \sqrt{a_N p^{N+1}} \prod_{j=1}^{N} \langle \boldsymbol{\omega}_j, \mathbf{x} \rangle$, where N is sampled with prob. $1/p^{N+1}$ for the parameter $p > 1$ and $\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_N$ are Rademacher distributed. Then,

$$k(\mathbf{x}, \mathbf{z}) = \sum_{i=0}^{\infty} \Psi_i(\mathbf{x})\Psi_i(\mathbf{z}) \approx \sum_{i=0}^{M} \Psi_i(\mathbf{x})\Psi_i(\mathbf{z})$$

where the approximation holds in mean over $\boldsymbol{\omega}_j$ and uniformly in concentration for $\mathbf{x}, \mathbf{z}$ lying in a compact set. $\Rightarrow$ `for the experiments` $M = 3n$

Kernelizing the Covariance Representation

For $i = 1, \dots, M$ we can `compute`[a] $\mathbf{h}_i$ such that

$$\mathbf{\Psi}|_M(\mathbf{h}_i) = \frac{1}{\sqrt{M}}[\Psi_1(\mathbf{h}_i), \dots, \Psi_M(\mathbf{h}_i)] = \mathbf{e}_i.$$

Therefore, once defined $K_{is}[\mathbf{X}, \mathbf{h}] = k(\mathbf{x}(s), \mathbf{h}_i)$,

$$\boxed{\widehat{\mathbb{S}}(k) = \mathbf{K}[\mathbf{X}, \mathbf{h}]\mathbf{P}\mathbf{K}[\mathbf{X}, \mathbf{h}]^\top} = \widehat{\mathbb{S}}(\mathbf{\Psi}|_M(\mathbf{X}))$$

$$\approx \widehat{\mathbb{S}}(\mathbf{\Psi}(\mathbf{X})) = \widehat{\mathbb{S}}(\mathbf{\Phi}(\mathbf{X})).$$

[a] see Proposition 1. in the paper

The Pseudocode

1. `For each action, extract the data matrix` $\mathbf{X}$ `collecting all the` T `temporal observations` $\mathbf{x}(1), \dots, \mathbf{x}(T)$`, each one encoding the 3D coordinates of the` n `joints.`
2. `For each data matrix` $\mathbf{X}$`, select` $\mathbf{h}_1, \dots, \mathbf{h}_M$ `as in Proposition 1. and compute` $\mathbf{K}[\mathbf{X}, \mathbf{h}]$.
3. `Compute the linear operator` $\mathbf{P}$.
4. `By means of` $\mathbf{K}[\mathbf{X}, \mathbf{h}]$ `and` $\mathbf{P}$`, compute` $\widehat{\mathbb{S}}(k)$.

Experimental Results

Publicly available code at `https://www.iit.it/pavis/code/kcar`

For any action trial $\mathbf{X}$, compute the kernel matrix $\mathbf{K}[\mathbf{X}, \mathbf{h}]$ through $k(\mathbf{x}(s), \mathbf{h}_i) = \exp(\gamma \cdot \langle \mathbf{x}(s), \mathbf{h}_i \rangle)$ for $s = 1, \dots, T$ and $i = 1, \dots, 3n$. The final classification step is performed with a support vector machine (SVM) fed with a log-Euclidean Gaussian kernel. The SVM cost value C and the kernel parameter γ are cross-validated.

Method	MSR-Action3D
Action Graph [4]	79.0%
Random Occupancy Patterns [8]	86.5%
Actionlets [9]	88.2%
Pose Set [7]	90.0%
Moving Pose [12]	91.7%
Lie Group [6]	92.5%
Normal Vectors [11]	93.1%
Kernelized-COV (proposed)	**96.2%**

Method	MSR-Action3D	MSR-Daily-Activity	MSRC-Kinect12	HDM-05
Region-COV [5]	74.0%	85.0%	89.2%	91.5%
Hierarchy of COVs [2]	90.5%	–	91.7%	–
COV-$J_{\mathcal{H}}$-SVM [1]	80.4%	75.5%	89.2%	82.5%
Ker-RP-POL [10]	96.2%	**96.9%**	90.5%	93.6%
Ker-RP-RBF [10]	**96.9%**	96.3%	92.3%	96.8%
Kernelized-COV (proposed)	96.2%	96.3%	**95.0%**	**98.1%**

[1] Mehrtash Harandi, Mathieu Salzmann, and Fatih Porikli. Bregman divergences for infinite dimensional covariance matrices. In *CVPR*, 2014.
[2] M. Hussein, M. Torki, M. Gowayyed, and M. El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. *IJCAI*, 2013.
[3] P. Kar and H. Karnick. Random features maps for dot product kernels. In *AISTATS*, 2012.
[4] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3D points. In *CVPRw*, 2010.
[5] Oncel Tuzel, Fatih Porikli, and Peter Meer. Region covariance: A fast descriptor for detection and classification. In *ECCV*. 2006.
[6] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3D skeletons as points in a lie group. In *CVPR*, 2014.
[7] Chunyu Wang, Yizhou Wang, and Alan L. Yuille. An approach to pose-based action recognition. In *CVPR*, 2013.
[8] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3D action recognition with random occupancy patterns. In *ECCV*, 2012.
[9] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012.
[10] Lei Wang, Jianjia Zhang, Luping Zhou, Chang Tang, and Wanqing Li. Beyond covariance: Feature representation with nonlinear kernel matrices. In *ICCV*, 2015.
[11] Xiaodong Yang and Yingli Tian. Super normal vector for activity recognition using depth sequences. In *CVPR*, 2014.
[12] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. In *ICCV*, 2013.