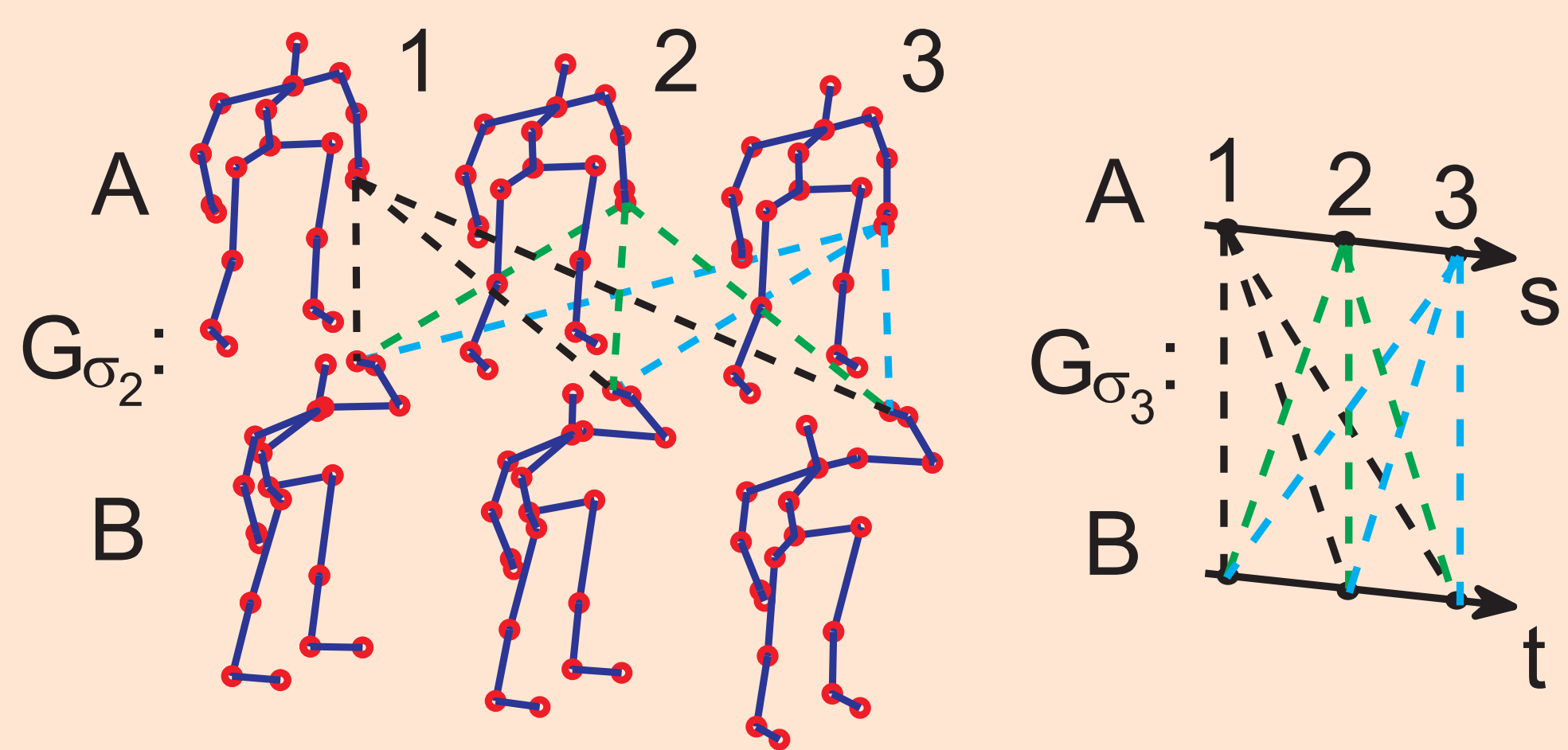


## Contributions and key ideas

- We define RBF kernels on 3D joint sequences that can compactly capture higher-order relationships between skeleton joints for 3D action recognition.
- We embed low-level 3D joints and time index into Reproducing Kernel Hilbert Space.
- We devise a *sequence compatibility kernel* that captures the spatio-temporal compatibility of joints in one sequence against those in the other sequence.
- We devise a *dynamics compatibility kernel* that explicitly models the action dynamics of a sequence, e.g. captures misplacement vectors between body joints across sequences.
- Evaluating kernels is costly. Thus, we linearize the kernels to form kernel descriptors. The higher-order outer-products derived from these kernel descriptors become higher-order tensor representations.
- Invariance to spatial and temporal variations is achieved by choosing the right kernel radii.

## Sequence Compatibility Kernel (SCK)



- We rewrite Gaussian between  $\mathbf{u} \in \mathbb{R}^{d'}$  and  $\bar{\mathbf{u}} \in \mathbb{R}^{d'}$ :

$$G_\sigma(\mathbf{u} - \bar{\mathbf{u}}) = e^{-\|\mathbf{u} - \bar{\mathbf{u}}\|_2^2 / 2\sigma^2} = \left(\frac{2}{\pi\sigma^2}\right)^{\frac{d'}{2}} \int_{\zeta \in \mathbb{R}^{d'}} G_{\sigma/\sqrt{2}}(\mathbf{u} - \zeta) G_{\sigma/\sqrt{2}}(\bar{\mathbf{u}} - \zeta) d\zeta.$$

- We use finite approximation by  $\zeta_1, \dots, \zeta_Z$  pivots:  
 $\phi(\mathbf{u}) = \left[ G_{\sigma/\sqrt{2}}(\mathbf{u} - \zeta_1), \dots, G_{\sigma/\sqrt{2}}(\mathbf{u} - \zeta_Z) \right]^T$   
 and  $G_\sigma(\mathbf{u} - \bar{\mathbf{u}}) \approx \langle \sqrt{c}\phi(\mathbf{u}), \sqrt{c}\phi(\bar{\mathbf{u}}) \rangle$ .

- We define a pose sequence  $\Pi = \{\mathbf{x}_{is} \in \mathbb{R}^3, i \in \mathcal{I}_J, s \in \mathcal{I}_N\}$ , where each pose consists of  $J$  body-keypoints, a sequence of  $N$  skeletons and  $\Pi$  is associated with one of  $K$  action class labels.
- Let  $\mathbf{x}_{is} \in \mathbb{R}^3$  and  $\mathbf{y}_{jt} \in \mathbb{R}^3$  correspond to the body-joint coordinates of two sequences  $\Pi_A$  and  $\Pi_B$ .
- We define our *sequence compatibility kernel* (SCK) between sequences  $\Pi_A$  and  $\Pi_B$  as:

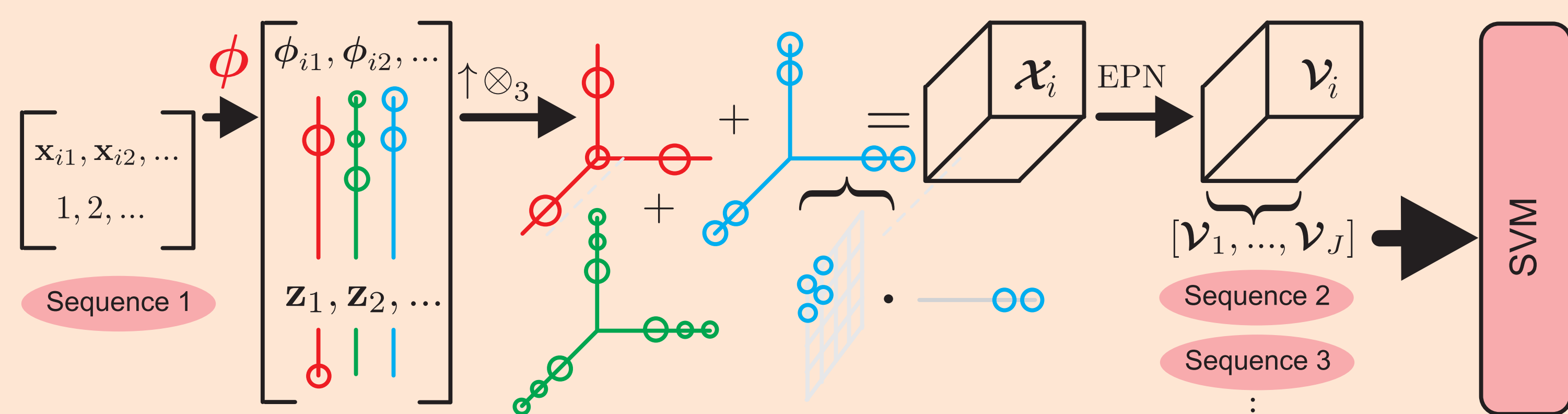
$$K_S(\Pi_A, \Pi_B) = \frac{1}{\Lambda} \sum_{(i,s) \in \mathcal{J}} \sum_{(j,t) \in \mathcal{J}} G_{\sigma_1}(i-j) \left( \beta_1 G_{\sigma_2}(\mathbf{x}_{is} - \mathbf{y}_{jt}) + \beta_2 G_{\sigma_3}\left(\frac{s-t}{N}\right) \right)^r,$$

where  $\Lambda$  is a normalization constant,  $\mathcal{J} = \mathcal{I}_J \times \mathcal{I}_N$  and  $r$  is the order of the statistics, e.g.  $r \geq 3$ .

- Kernels  $G_{\sigma_1}, G_{\sigma_2}, G_{\sigma_3}$  capture the compatibility between (i) joint-types  $i$  and  $j$ , (ii) joint locations  $\mathbf{x}$  and  $\mathbf{y}$ , (iii) the temporal alignment of two poses.
- $\beta_1, \beta_2 \geq 0$  adjust the importance of the body-joint compatibility against the temporal alignment.

Then, we obtain a linearized representation  $K_S^*(\Pi_A, \Pi_B) = \langle \mathbf{V}_A, \mathbf{V}_B \rangle$ , where:

$$\mathbf{V} = [\mathcal{G}(\mathcal{X}_i)]_{i \in \mathcal{I}_J}^{\oplus 4} \text{ and } \mathcal{X}_i = \frac{1}{\sqrt{\Lambda}} \sum_{s \in \mathcal{I}_N}^{\uparrow \otimes r} \begin{bmatrix} \sqrt{\beta_1} \phi(\mathbf{x}_{is}) \\ \sqrt{\beta_2} \mathbf{z}(s/N) \end{bmatrix}$$



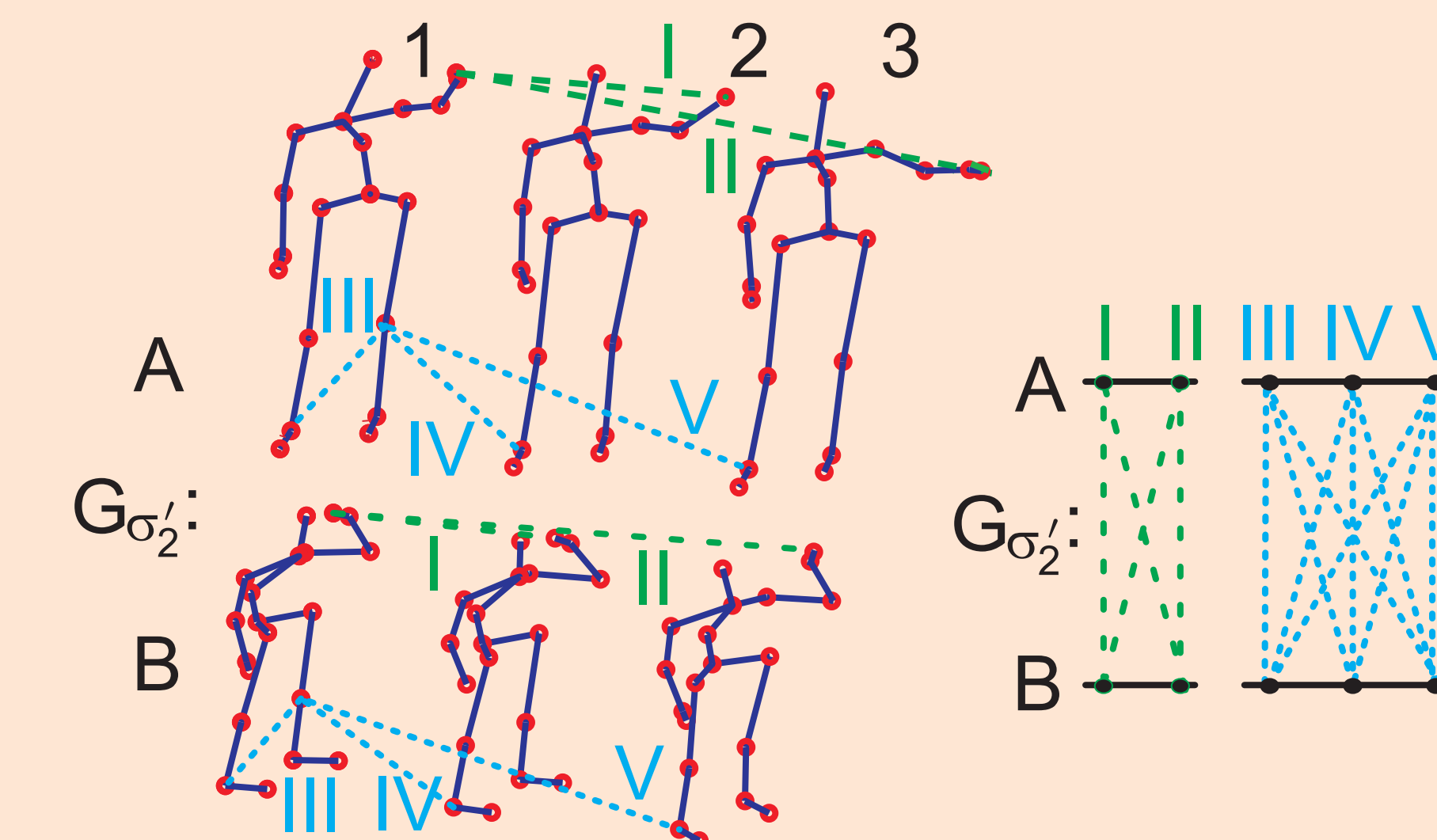
## Dynamics Compatibility Kernel (DCK)

- The intra-sequence spatio-temporal dynamics is captured for any two action sequences  $\Pi_A$  and  $\Pi_B$  by:

$$K_D(\Pi_A, \Pi_B) = \frac{1}{\Lambda} \sum_{\substack{(i,s) \in \mathcal{J}, \\ (i',s') \in \mathcal{J}, \\ i \neq i', s' \neq s}} \sum_{\substack{(j,t) \in \mathcal{J}, \\ (j',t') \in \mathcal{J}, \\ j \neq j', t' \neq t}} G'_{\sigma'_1}(i-j, i'-j') G_{\sigma'_2}((\mathbf{x}_{is} - \mathbf{x}_{i's'}) - (\mathbf{y}_{jt} - \mathbf{y}_{j't'})) G'_{\sigma'_3}\left(\frac{s-t}{N}, \frac{s'-t'}{N}\right) G'_{\sigma'_4}(s-s', t-t'),$$

where  $G'_\sigma(\alpha, \beta) = G_\sigma(\alpha)G_\sigma(\beta)$ .

- Kernel  $G'_{\sigma'_1}$  captures sensor uncertainty in body-keypoint detection (we use a delta function).
- Kernel  $G_{\sigma'_2}$  models the spatio-temporal co-occurrences of the body-joints.
- Kernels  $G_{\sigma'_3}$  encode the temporal start and end-points  $(s, s')$  from  $\Pi_A$  and  $(t, t')$  from  $\Pi_B$ .
- Kernels  $G_{\sigma'_4}$  limits contributions of dynamics between temporal points ( $\sigma'_4$  is small) if they are distant from each other, i.e. if  $s' \gg s$  or  $t' \gg t$ .



- Our final representation can be expressed as follows:

$$K_D^*(\Pi_A, \Pi_B) = \langle \mathbf{V}_A, \mathbf{V}_B \rangle \text{ and } \mathbf{V} = [\mathbf{v}_{ii'}]_{i>i', i, i' \in \mathcal{I}_J}^{\oplus 4}, \mathbf{v}_{ii'} = \mathcal{G}(\mathcal{X}_{ii'})$$

and  $\mathcal{X}_{ii'} = \frac{1}{\sqrt{\Lambda}} \sum_{\substack{s \in \mathcal{I}_N, \\ s' \in \mathcal{I}_N, \\ s \neq s'}} G_{\sigma'_4}(s-s') \left( \phi(\mathbf{x}_{is} - \mathbf{x}_{i's'}) \cdot \mathbf{z}\left(\frac{s}{N}\right)^T \right) \uparrow \otimes \mathbf{z}\left(\frac{s'}{N}\right).$

- Eigenvalue Power Normalisation  $\mathcal{G}$  deals with correlated bursts in each sequence:

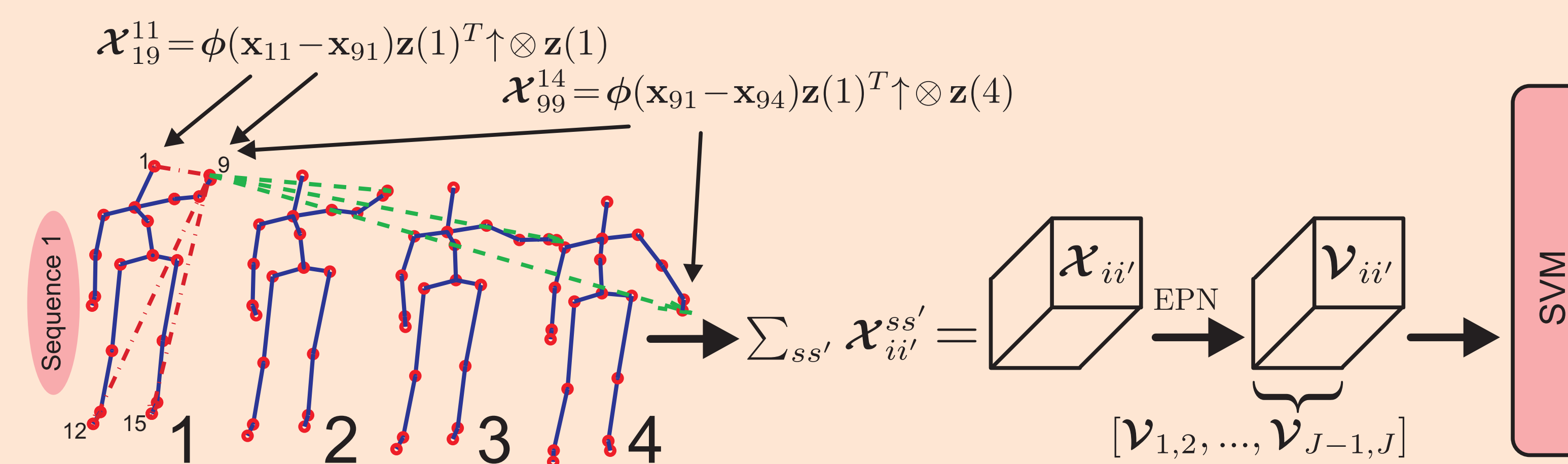
$$(\mathcal{E}; \mathbf{A}_1, \dots, \mathbf{A}_r) = \text{HOSVD}(\mathcal{X})$$

$$\hat{\mathcal{E}} = \text{Sgn}(\mathcal{E}) |\mathcal{E}|^\gamma$$

$$\hat{\mathcal{X}} = ((\hat{\mathcal{E}} \otimes_1 \mathbf{A}_1) \dots) \otimes_r \mathbf{A}_r$$

$$\mathbf{V} = \mathcal{G}(\mathcal{X}) = \text{Sgn}(\hat{\mathcal{X}}) |\hat{\mathcal{X}}|^{\gamma^*}$$

- We illustrate below the captured statistics:



## Results

- Florence3D-Action dataset:

	SCK	DCK	SCK+DCK
accuracy	92.98%	93.03%	92.77%
Bag-of-Poses	82.00% [4]	SE(3) 90.88% [5]	

- UTKinect-Action dataset:

	SCK	DCK	SCK+DCK
accuracy	96.08%	97.5%	98.2%
3D joints. hist.	90.92% [6]	SE(3) 97.08% [5]	

- MSR-Action3D dataset (SCK only):

order	slice-wise EPN	HOSVD EPN	SCK+DCK
order $r=3$	90.72%	94.02%	SE(3) 89.48%
order $r=3$			
order $r=4$			95.2%

- [1] Higher-order Occurrence Pooling for Bags-of-Words: Visual Concept Detection. P. Koniusz, F. Yan, P. H. Gosselin, K. Mikolajczyk. TPAMI 2016.
- [2] Sparse Coding for Third-order Super-symmetric Tensor Descriptors with Application to Texture Recognition. P. Koniusz, A. Chelian. CVPR 2016.
- [3] Convolutional Kernel Networks. J. Mairal, P. Koniusz, Z. Harchaoui, C. Schmid. NIPS 2014.
- [4] Recognizing Actions from Depth Cameras as Weakly Aligned Multi-part bag-of-poses. L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, P. Pala. CVPR Workshops 2013.
- [5] Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group. R. Vemulapalli, F. Arrate, R. Chellappa. CVPR 2014.
- [6] View Invariant Human Action Recognition Using Histograms of 3D Joints. L. Xia, C. Chen, J. K. Aggarwal. CVPR Workshops 2012.