A time series kernel for action recognition

Adrien Gaidon
http://lear.inrialpes.fr/people/gaidon
Zaid Harchaoui
http://lear.inrialpes.fr/people/harchaoui
Cordelia Schmid
http://lear.inrialpes.fr/people/schmid

LEAR - INRIA Grenoble, LJK
655, avenue de l'Europe
38330 Montbonnot, France

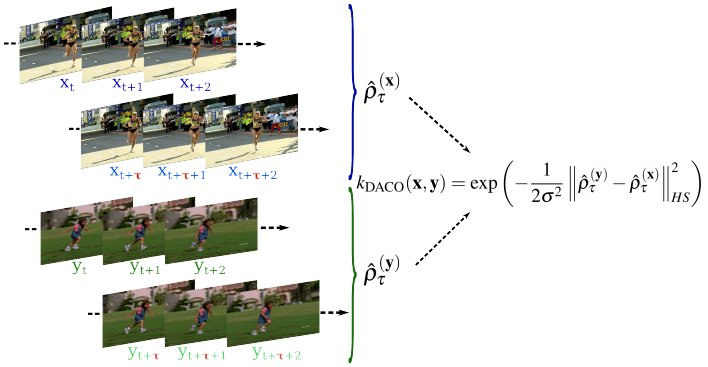

Figure 1: Computation of our DACO kernel. For two actions represented as time series of frames, $\mathbf{x} = (x_1, \ldots, x_T)$ and $\mathbf{y} = (y_1, \ldots, y_{T'})$, the kernel compares their dynamics by using the difference between their auto-correlations $\hat{\rho}_\tau^{(\mathbf{x})}$ and $\hat{\rho}_\tau^{(\mathbf{y})}$, with a lag of τ frames.

We address the problem of supervised action recognition, *i.e.* deciding whether an action is performed in a video or not. Previous approaches on realistic videos generally focus on models aggregating statistics of local features over the entire duration of an action (e.g. in a bag-of-features, BOF, model [2]), thus discarding temporal information. Recent efforts [1, 4] have been made to improve these models by incorporating global temporal information but are not tailored to capture temporal relationships between frames. Therefore, we provide a new kernel to compare the dynamics and temporal structure of actions, directly represented as time series of frames.

Our kernel between two actions is based on the distance between their respective auto-correlations. The auto-correlation

$$\hat{\rho}_\tau^{(\mathbf{x})} = \left(\hat{\Sigma}^{(\mathbf{x})} + \gamma \mathbf{I}\right)^{-1} \hat{\Sigma}_\tau^{(\mathbf{x})}$$

at time-lag τ for action $\mathbf{x} = (x_t)_{t=1\cdots T}$ is the cross-covariance $\hat{\Sigma}_\tau^{(\mathbf{x})}$ between the time series and a version shifted by τ frames, normalized by the regularized inverse of the covariance $\hat{\Sigma}^{(\mathbf{x})}$. Auto-correlation contains information pertaining to the temporal dependencies between frames and the temporal structure of actions, as it depends on the ordering of the frames.

Hence, we propose to compare two actions $\mathbf{x}$ and $\mathbf{y}$ by using the Hilbert-Schmidt norm of the difference between auto-correlations and we call the associated Gaussian RBF kernel the *Difference between Auto-Correlation Operators* (DACO) kernel:

$$k_{\mathrm{DACO}}(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{d_{\mathrm{DACO}}(\mathbf{x},\mathbf{y})^2}{2\sigma^2}\right), \; d_{\mathrm{DACO}}(\mathbf{x},\mathbf{y}) = \left\|\hat{\rho}_\tau^{(\mathbf{y})} - \hat{\rho}_\tau^{(\mathbf{x})}\right\|_{HS}$$

The Hilbert-Schmidt norm is defined as $\|A\|_{HS}^2 = \mathrm{Tr}\,(A^* A)$.

Frame representations are in general high-dimensional (*e.g.* several thousands of dimensions for BOF) and might be of a non-vector type. Therefore, instead of making assumptions on the frame models, we only assume the availability of a symmetric positive-definite kernel between frames $k_F : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$, such as the intersection the kernel between per-frame BOF. Consequently, we derive the formula of our DACO action kernel in the feature space induced by a frame kernel, which only involves the between-frames Gram matrix $\mathbf{K}$:

$$d_{\mathrm{DACO}}(\mathbf{x},\mathbf{y})^2 \;=\; \mathrm{Tr}\left(\mathbf{N}^T \mathbf{K} \mathbf{N} \mathbf{K}_{+\tau}\right)$$

where $\mathbf{N}^T$ is also expressed in function of $\mathbf{K}$. In our experiments, we chose to use the state-of-the-art MBH features recently proposed by Wang *et al.* [7] and represent each frame by a BOF. We use the intersection kernel between histograms as base kernel between frames.

	KTH	UCF Sports	Youtube
Wang MBH [7]	**95.0**	84.8	83.9
DME	94.8	87.0	86.7
DACO	93.4	85.3	79.1
DME+DACO	**94.9**	**90.3**	**87.9**

Table 1: Average accuracy on KTH, UCF Sports and Youtube.

Dynamic aspects alone are likely to be insufficient to describe some types of actions, especially when context or the nature of objects involved is discriminative. Therefore, our goal is to show that our DACO kernel is complementary with orderless aggregation statistics. We provide results for a kernel that is the linear combination of our DACO kernel and the *Difference between Mean Elements* (DME) kernel:

$$k_{DME}(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{1}{2\sigma^2} d_{DME}(\mathbf{x},\mathbf{y})^2\right) \quad , \quad d_{DME}(\mathbf{x},\mathbf{y}) = \left\|\hat{\mu}_y - \hat{\mu}_x\right\|_{\mathscr{H}_F}$$

This kernel is simply using the difference between the means, $\hat{\mu}_x$ and $\hat{\mu}_y$, of the frames in the feature space $\mathscr{H}_F$ induced by the kernel on frames. This is related to the traditional BOF approach consisting of aggregating local descriptors computed over the entire video sequence like in [2, 7], except that the aggregation is performed in the feature space. In practice, the DME kernel is computed using the same kernel matrix $\mathbf{K}$ as DACO:

$$d_{DME}(\mathbf{x},\mathbf{y})^2 = \mathbf{m}^T \mathbf{K} \mathbf{m} \quad , \quad \mathbf{m} = \Big[\underbrace{-\frac{1}{T},\cdots,-\frac{1}{T}}_{T}, \underbrace{\frac{1}{T'},\cdots,\frac{1}{T'}}_{T'}\Big]^T$$

We investigate the use of our kernel on three state-of-the-art publicly available datasets: the **KTH** [6], **UCF Sports** [5] and **Youtube** [3] datasets. We compare average classification accuracies in table 1. We achieve similar or better performance than the state of the art with the simple combination of the aggregation-based DME kernel and our auto-correlation-based DACO kernel. Furthermore, the combination is always superior to the best one of the two. This is true even in the case of the Youtube dataset, where the DACO kernel performs clearly worse than DME. This shows that DACO is more suited to short duration, fast actions and is explained by: (i) the small value of the time-lag we use ($\tau = 1$ frame) and (ii) long range temporal dependencies between frames of non-periodic actions are difficult to estimate. This suggests a possible improvement by applying our DACO kernel in a more temporally localized manner, in order to detect correlated components with a strong temporal structure.

[1] A. Gaidon, Z. Harchaoui, and C. Schmid. Actom sequence models for efficient action detection. In *CVPR*, 2011.

[2] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[3] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *CVPR*, 2009.

[4] J.C. Niebles, C. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010.

[5] M.D. Rodriguez, J. Ahmed, and M. Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.

[6] C. Schueldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, 2004.

[7] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action recognition by dense trajectories. In *CVPR*, 2011.