Adaptive Structured Pooling for Action Recognition

Svebor Karaman[1]
svebor.karaman@unifi.it
Lorenzo Seidenari[1]
lorenzo.seidenari@unifi.it
Shugao Ma[2]
shugaoma@bu.edu
Alberto Del Bimbo[1]
alberto.delbimbo@unifi.it
Stan Sclaroff[2]
sclaroff@bu.edu

[1] MICC (Media Integration and Communication Center)
University of Florence
Florence, Italy

[2] Boston University
Boston, USA

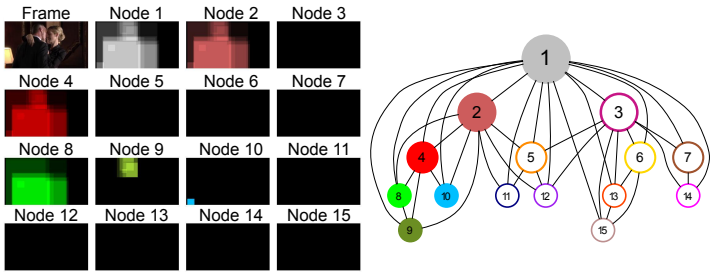

Figure 1: Overview of our method. Left: a frame of the "kiss" action of the HighFive dataset and pooling maps plots. Right: the video structure graph where nodes are spatio-temporal pooling regions at different granularities. Note how node 9 selects both actors faces.

We propose an adaptive structured pooling strategy to solve the action recognition problem in videos. Our method aims at individuating several spatio-temporal pooling regions each corresponding to a consistent spatial and temporal subset of the video. Each of them gives a pooling weight map and is represented as a Fisher vector computed from the soft weighted contributions of all dense trajectories evolving in it. We further represent each video through a graph structure, defined over multiple granularities of spatio-temporal subsets. The graph structures extracted from all videos are compared with an efficient graph matching kernel.

Soft pooling weigths. Given a set of Hierarchical Space-Time Segments (HSTS) [2] $\mathcal{S}_k$ we define a weighted pooling map M_k by accumulating how many segments of $\mathcal{S}_k$ are present in each frame at each position. For every pixel $p=(x,y)$ of frame t, we compute the pooling map value $M_k^t(p)$ as the count of segment enclosing this position $M_k^t = \sum_{s\in\mathcal{S}_k^t} \Psi_s$ where for each segment $s \in \mathcal{S}_k^t$ we define the function $\Psi_s(p)=1$ if $p \in s$ and $\Psi_s(p)=0$ otherwise.

The pooling map M_k^t is further normalized by the total number of segments in the frame and square-rooted. This pooling maps represent at any moment of the video, how much each pixel is relevant with respect to the set $\mathcal{S}_k$. The more segments overlap in one position the more likely this pixel is significant for the action taking place. Finally, for a video with T frames we define the spatio-temporal pooling map as:

$$M_k(x,y,t) = \left\{ M_k^1(x,y) \dots M_k^T(x,y) \right\} \tag{1}$$

For each local feature to be encoded, we estimate the weight with respect to to set $\mathcal{S}_k$ as a small local integral of the pooling map M_k around its centroid. That is for each $x_m \in X$ with the spatio-temporal coordinates of its centroid being $(x_{x_m}, y_{x_m}, t_{x_m})$, w_m^k is estimated as:

$$w_m^k = \int_{x_{x_m}-v_x}^{x_{x_m}+v_x} \int_{y_{x_m}-v_y}^{y_{x_m}+v_y} \int_{t_{x_m}-v_t}^{t_{x_m}+v_t} M_k(x,y,t)\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}t \tag{2}$$

Finally, all weights of a pooling region are normalized to sum to one in order to have comparable representation no matter how many number of features are present in the region. We obtain soft-pooling by using the weight w_m^k of each feature $x_m \in X$ within the soft Fisher encoding formulation (see eq. 3 and 4).

Fisher encoding with soft pooling. Given the Gaussian Mixture Model (GMM) $u_\lambda = \sum_{n=1}^N \omega_n u_n(x;\mu_n,\sigma_n)$ and the M features of X, we compute for each component u_n the mean $\mathcal{G}_n^\mu(X)$ and covariance elements $\mathcal{G}_n^\sigma(X)$ of a Fisher vector as:

$$\mathcal{G}_n^\mu(X) = \frac{1}{\sqrt{\omega_n}} \sum_{m=1}^M w_m \gamma_n(x_m) \left(\frac{x_m-\mu_n}{\sigma_n} \right), \tag{3}$$

$$\mathcal{G}_n^\sigma(X) = \frac{1}{\sqrt{2\omega_n}} \sum_{m=1}^M w_m \gamma_n(x_m) \left(\frac{(x_m-\mu_n)^2}{\sigma_n^2} - 1 \right), \tag{4}$$

where $\gamma_n(x_m)$ is the posterior probability of the feature x_m for the component n of the GMM and w_m is the weight obtained from eq. 2.

Spatio-temporally structured pooling of a video. We want to build a structured representation of each video. We propose to find coherent subsets by grouping together segments according to their overlap. This will create a set of local (both spatially and temporally) pooling regions.

We first compute an affinity matrix A of all segments $\mathcal{S}$ of a video. The affinity of two segments s_i (alive from frame t_{is} to t_{ie}) and s_j (alive from frame t_{js} to t_{je}) is computed as:

$$A(s_i,s_j) = \frac{1}{\min(t_{ie}-t_{is}, t_{je}-t_{js})} \sum_{t\in[\max(t_{is},t_{js}),\min(t_{ie},t_{je})]} \frac{s_i^t \cap s_j^t}{s_i^t \cup s_j^t}. \tag{5}$$

Given this affinity matrix we run the normalized cuts algorithm to obtain the subsets of segments. Instead of choosing one fixed number of subsets, we use multiple increasing sizes that will each provide a set of finer local representations of the video. We represent each HSTS cluster as a node in the graph, and each node attribute is the soft pooling of dense trajectories features weighted by the map computed on all segments of this cluster. We link clusters based on their overlap, we create a link between all clusters that have at least a pair of overlapping segments (even partially). An illustration of one video graph is shown in Figure 1. To compare the video graphs we use the efficient GraphHopper kernel from [1].

Conclusions. Our structured representation is adaptive to the content of the video and does not rely on a fixed partition of neither space nor time. We exploit an unsupervised procedure to generate a structured representation of the video. Our representation jointly models the hierarchical and spatio-temporal relationship of videos without imposing a strict hierarchy.

Experiments conducted on two standard datasets for action recognition show a significant improvement over the state-of-the-art. We obtain **65.4%** mean AP on HighFive dataset and **90.4%** mean per class accuracy on UCF Sports dataset. In the future, we would like to see if our structured representation could also be used to solve the action localization problem by identifying the paths and/or nodes that are most relevant for the action.

[1] Aasa Feragen, Niklas Kasenburg, Jens Petersen, Marleen de Bruijne, and Karsten Borgwardt. Scalable kernels for graphs with continuous attributes. In *Advances in Neural Information Processing Systems*, pages 216–224, 2013.

[2] Shugao Ma, Jianming Zhang, Nazli Ikizler-Cinbis, and Stan Sclaroff. Action recognition and localization by hierarchical space-time segments. In *Proc. of International Conference on Computer Vision (ICCV)*. IEEE, 2013.