Adaptive Structured Pooling for Action Recognition

Svebor Karaman, **Lorenzo Seidenari**, Shugao Ma, Alberto Del Bimbo, Stan Sclaroff

{svebor.karaman,lorenzo.seidenari,alberto.delbimbo}@unifi.it
{shugaoma, sclaroff}@bu.edu

MICC Media Integration and Communication Center

BOSTON UNIVERSITY

Objectives

What do we want to do?

- Find coherent spatio-temporal regions in video
- Build a structured representation of a video

How do we do it?

- Cluster HSTS [4] according to their overlap
- Define soft pooling regions from each cluster
- Link overlapping clusters in a graph structure
- Graphs comparison with GraphHopper kernel [1]

HSTS

- Hierarchical Space-Time Segments (HSTS) represent spatio-temporal dynamic portions of a video preserving both moving and static relevant regions
- Obtained by a hierarchical segmentation relying both on image and motion and short term tracking, see [4]

Soft Pooling Weights

Weighted pooling map M_k by accumulating, in each frame t at each position p, segments of a HSTS set $\mathcal{S}_k$:

$$M_k^t = \sum_{s \in \mathcal{S}_k^t} \Psi_s \, ,$$

where $s \in \mathcal{S}_k^t$ and $\Psi_s(p) = 1$ if $p \in s$ and $\Psi_s(p) = 0$ otherwise. M_k^t is L1-normalized and square-rooted. For a video of T frames:

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

For each feature $x_m \in X$, with $(x_{x_m}, y_{x_m}, t_{x_m})$ as spatio-temporal coordinates of its centroid, weight w_m^k as a local integral of the pooling map M_k:

$$w_m^k = \int_{x_{x_m}-v_x}^{x_{x_m}+v_x} \int_{y_{x_m}-v_y}^{y_{x_m}+v_y} \int_{t_{x_m}-v_t}^{t_{x_m}+v_t} M_k(x,y,t) \; dx \; dy \; dt$$

Soft Fisher Encoding

Given the GMM $u_\lambda = \sum_{n=1}^{N} \omega_n u_n(x; \mu_n, \sigma_n)$ and the M features of X, mean $\mathcal{G}_n^\mu(X)$ and covariance elements $\mathcal{G}_n^\sigma(X)$ of a Fisher vector for each u_n:

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

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

where $\gamma_n(x_m)$ is the posterior probability of the feature x_m for the component n of the GMM.

ST Structured Pooling

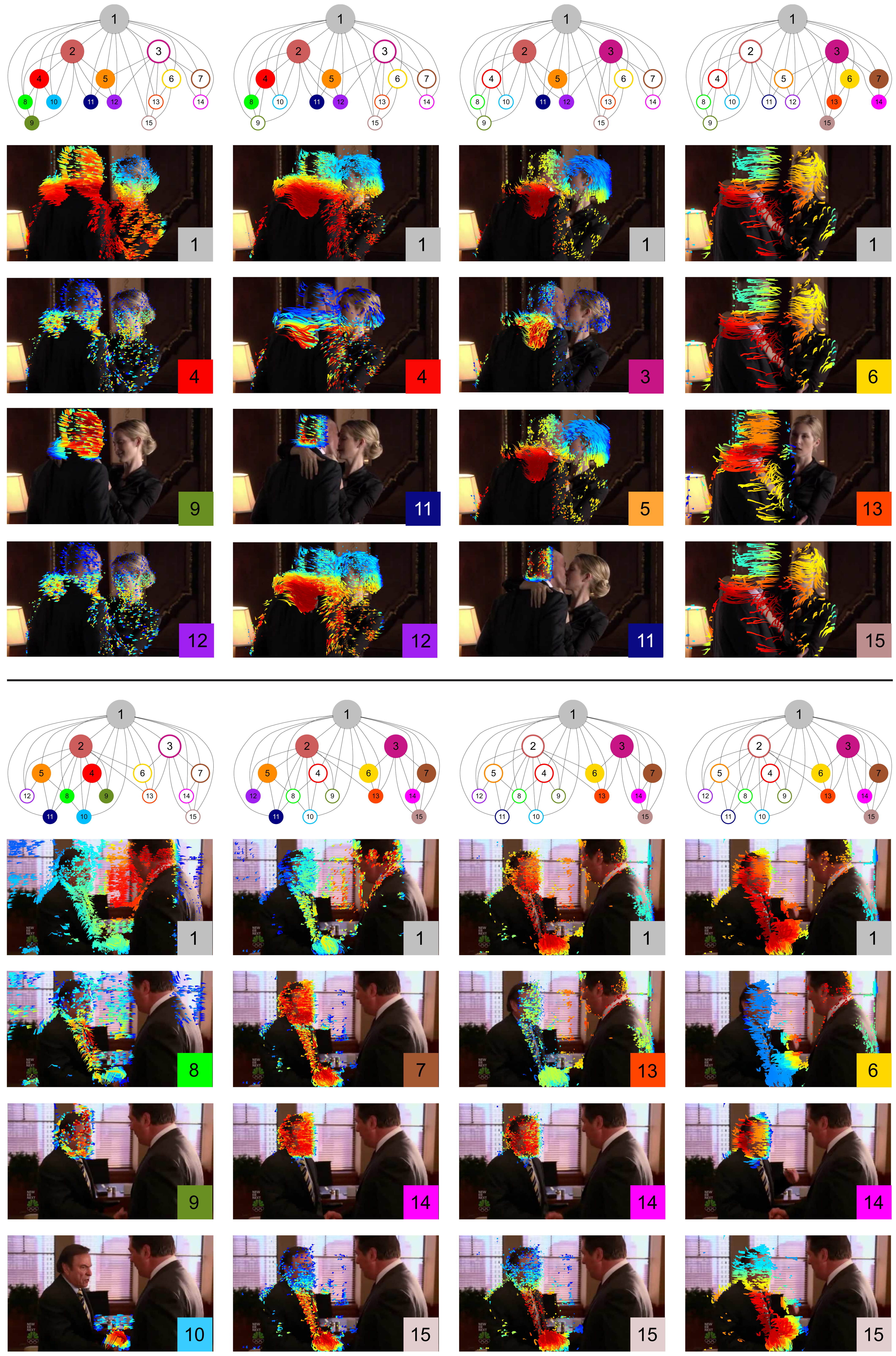
The affinity $A(s_i, s_j)$ of two segments s_i (alive from t_{is} to t_{ie}) and s_j (alive from t_{js} to t_{je}) is:

$$A(s_i, s_j) = \frac{1}{min_l} \sum_{t \in [max_s, min_e]} \frac{s_i^t \cap s_j^t}{s_i^t \cup s_j^t} .$$

where $min_l = \min(t_{ie} - t_{is}, t_{je} - t_{js})$, $max_s = \max(t_{is}, t_{js})$ and $min_e = \min(t_{ie}, t_{je})$.

- Multiple runs of normalized cut on affinity matrix A to obtain different number of segments clusters
- Each cluster is a node in the graph, soft pooling of dense trajectories features as attribute
- Link between clusters having overlapping segments

Adaptive spatio-temporal structured representation of video



Weighted trajectories for exemplar active nodes on a subset of frames for action “kiss” and “handshake”.

Results

UCF Sports

Method	LOO	Split [3]
Our (all features)	**90.4**	**90.8**
Our (MBH only)	88.3	90.0
FVB (all features)	88.6	89.4
Lan *et al.*	83.7	73.1
Kovashka *et al.*	87.3	-
Klaser *et al.*	86.7	-
Wang *et al.*	-	85.2
Ma *et al.* [4]	-	81.7

HighFive

Method	mAP
Our (all features)	**65.4**
Our (MBH only)	62.8
FVB (all features)	61.3
Gaidon *et al.* [2]	62.4
Ma *et al.* [4]	53.3
Wang *et al.*	53.4
Laptev *et al.*	36.9
Patron-Perez *et al.*	42.4

Conclusions

- Structured representation adaptive to video content that does not rely on fixed partition of space or time
- Unsupervised procedure to generate a structured representation of the video
- Jointly models the hierarchical and spatio-temporal relationship without imposing a strict hierarchy
- Significant improvement over the state-of-the-art on two standard datasets

References

[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] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. Activity representation with motion hierarchies. *International Journal of Computer Vision*, pages 1–20, 2013.

[3] Tian Lan, Yang Wang, and Greg Mori. Discriminative figure-centric models for joint action localization and recognition. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 2003–2010. IEEE, 2011.

[4] 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.