Training Deformable Object Models for Human Detection based on Alignment and Clustering

Benjamin Drayer and Thomas Brox

Computer Vision Group, University of Freiburg, Germany

Contributions

- Clustering based on pairwise distances that optimally align the samples
- More accurate part placement for object detection

Non-rigid Alignment

- Alignment in feature (HOG) space [3]
- Minimize:

$$E(\mathrm{u}) = E_D(\mathrm{u}) + E_P(\mathrm{u})$$

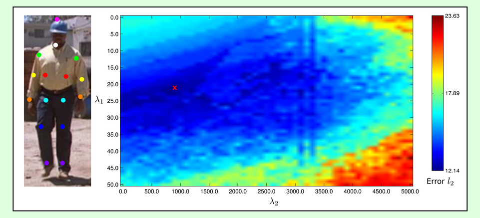
- Matching Cost:

$$E_D(\mathrm{u}) = \sum_{\mathrm{x}} \lambda_1 |F_2(\mathrm{x}+\mathrm{u}(\mathrm{x})) - F_1(\mathrm{x})|_1 - \lambda_2 \langle F_2(\mathrm{x}+\mathrm{u}(\mathrm{x})), F_1(\mathrm{x})\rangle$$

- Deformation Cost:

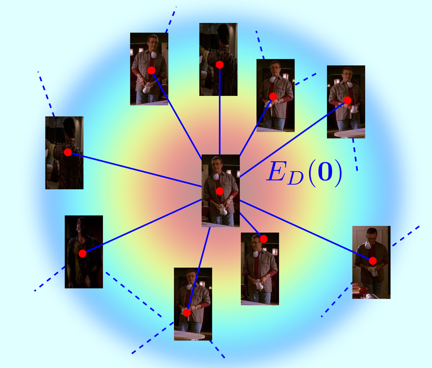
$$E_P(\mathrm{u}) = \sum_{\mathrm{x},\mathrm{y}\in\mathcal{N}(\mathrm{x})} |\mathrm{u}(\mathrm{x}) - \mathrm{u}(\mathrm{y})|_1 ,$$

- Optimize parameters λ_1, λ_2:

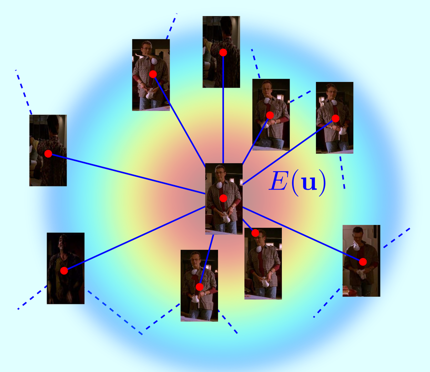


Clustering

Without alignment:



With alignment:



Affinities:

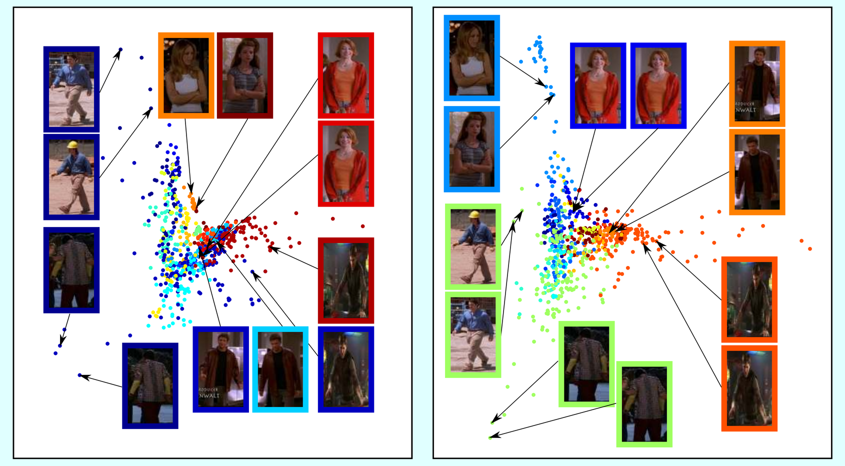
- Energy after alignment used as distance
- k-nearest neighbors

$$A(i,j) = \exp\left(-\frac{E(i,j)}{2\sigma^2}\right)$$

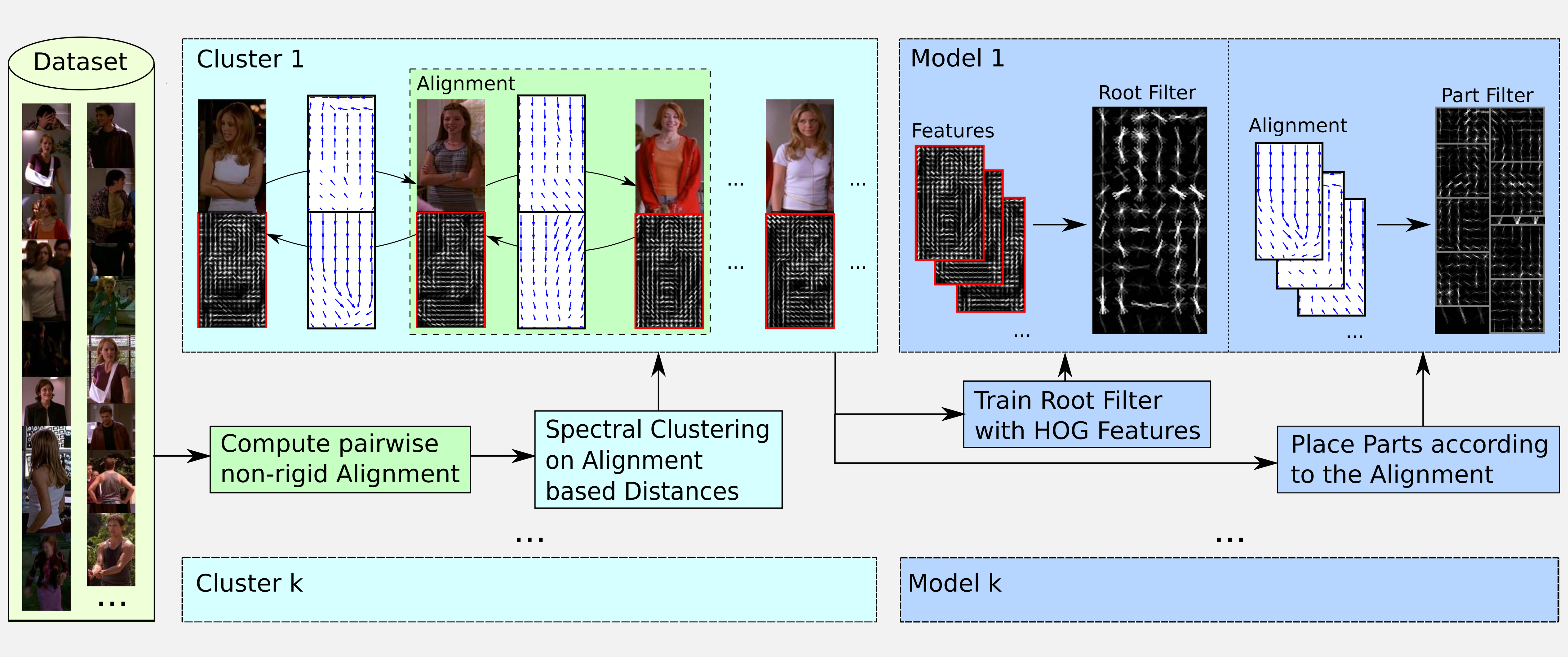
Clustering:

- Spectral clustering on $A + A^t$

Without alignment: With alignment:

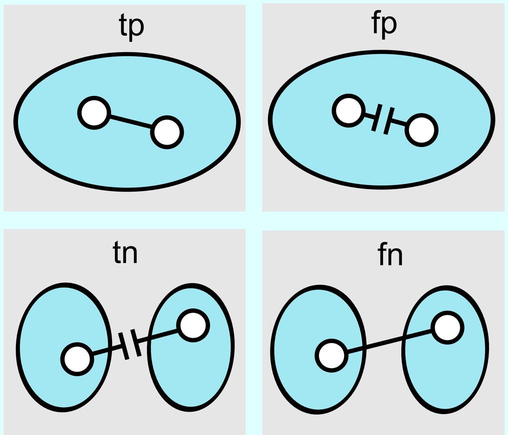


Approach Overview



Clustering Evaluation

- Evaluation based on average precision (AP) and F-measure
- Manually generated ground truth by labeling $m \approx 4000$ pairs



- Each pair is either:
  - Similar $\Rightarrow$ same cluster
  - Different $\Rightarrow$ different clusters
  - Ambiguous $\Rightarrow$ unknown
- No need to specify the number of clusters
- Due to random sampling and the effect of large numbers our deviation $\varepsilon \leq 1.6\%$

$$\varepsilon = z_{(1-\frac{\alpha}{2})} \frac{s}{\sqrt{m}}$$

- $z_{(1-\frac{\alpha}{2})}$ z-quantil of normal distribution
- $s \leq 0.5$ upper bound of standard deviation

Clustering Results

Without alignment	Energy	$(E_D(0))$	43.62
With alignment	Matching cost	$(E_D(\mathrm{u}))$	47.88
	Deformation cost	$(E_P(\mathrm{u}))$	19.72
	Both	$(E(\mathrm{u}))$	**48.04**

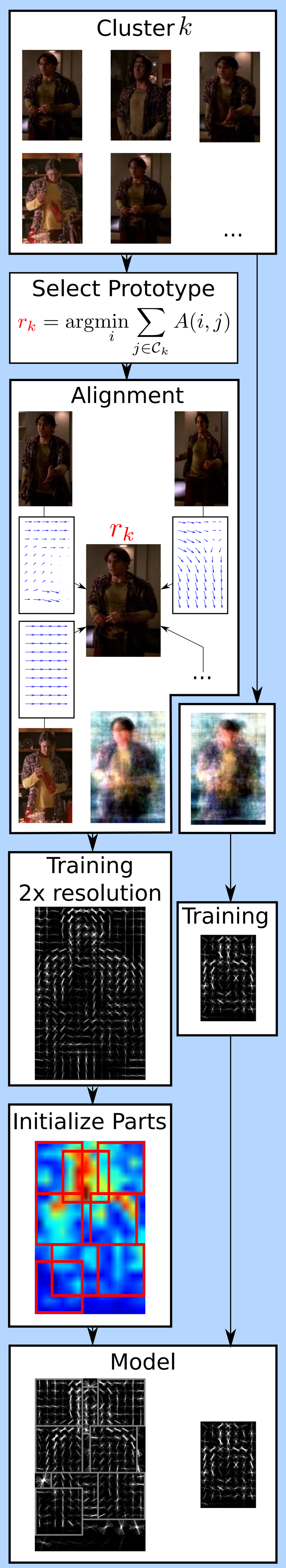
- Alignment based distances improve clustering performance (AP)

	Initialization	After DPM training
DPM $K=6$	0.3831	0.5012
DPM $K=10$	0.4664	0.5013
$E(\mathrm{u})$	0.5251	**0.5308**

- Our clustering works better than aspect ratio clustering, and it improves the cluster assignment during DPM training (F-measure)

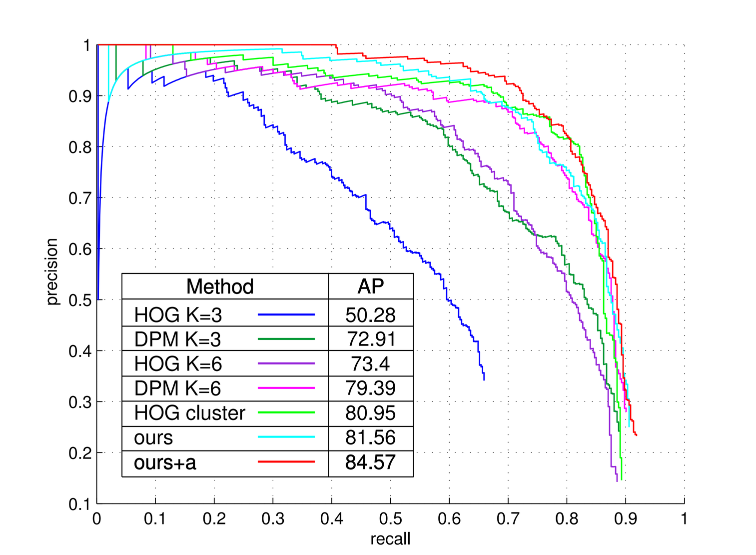
- Mean image with (right) and without (left) alignment

Training



Detection Results on Buffy Dataset

- Average precision on detection for various approaches on the Buffy dataset
- TP, if intersection over union ≥ 0.5
- Clustering and alignment improve AP by more than 5%



HOG $k=3$	DPM $k=3$	HOG $k=6$	DPM $k=6$	DPM $k=10$	HOG clustering	Ladicky [2]	DPM+c	DPM+a
50.28	72.91	73.4	79.39	78.04	80.95	76.03	81.56	84.57

Detection Results on PASCAL VOC

- Evaluation on the PASCAL VOC 2007 test set

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow
DPM [1]	28.9	59.5	10.0	15.2	25.5	49.6	57.9	19.3	22.4	25.2
DPM+c	29.7	58.2	9.7	16.3	22.9	50.3	52	14.8	18.9	27.9
DPM+a	33.2	57.4	9.7	16.9	25.0	48.6	52.3	13.3	20.2	30.3

	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
DPM [1]	23.3	11.1	56.8	48.7	41.9	12.2	17.8	33.6	45.1	41.6
DPM+c	24.9	10.3	57.2	48.7	36.8	12.9	17	24.1	45.8	40.9
DPM+a	26.6	6.5	60.1	49.1	38.4	9.8	18.7	29.7	47.3	39.8

- Classes with less variation or a denser sampling e.g. aeroplane improve with the alignment
- Classes with larger variability, such as cat, are hard to align and performance drops

This study was supported by the Excellence Initiative of the German Federal and State Governments (EXC 294) and by the ERC Starting Grant VIDEOLEARN.

[1] Felzenszwalb, P.F., Girshick, R.B., Mc Allester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. TPAMI (2010)

[2] Ladicky, L., Torr, P.H.S., Zisserman, A.: Latent svms for human detection with a locally affine deformation field. BMVC (2012)

[3] Drayer, B., Brox, T.: Distances based on non-rigid alignment for comparison of different object instances. GCPR (2013)