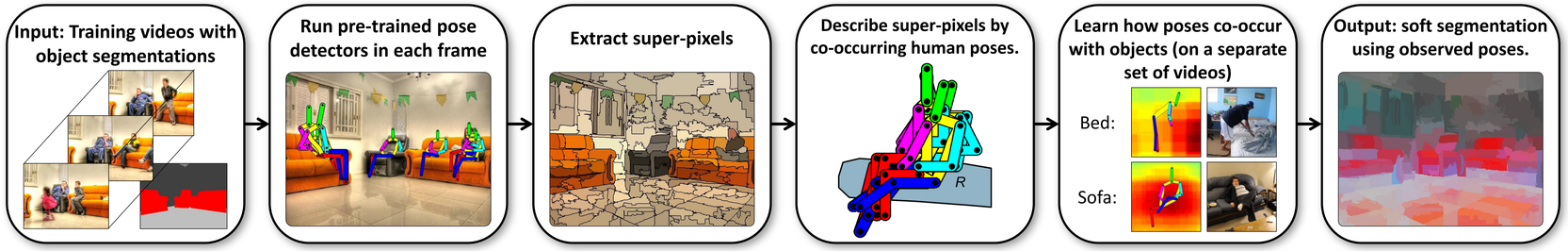
Scene semantics from long-term observation of people

Vincent Delaitre David F. Fouhey Ivan Laptev Josef Sivic Abhinav Gupta Alexei A. Efros

INRIA – WILLOW / École Normale Supérieure /Carnegie Mellon University

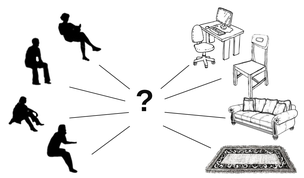


1 – Contribution

Goal: Recognize objects in realistic indoor scenes.

Motivation:

- Exploit the link between pose, action and object function.
- Use people in videos as active sensors to reason about the surrounding scene.



Contributions:

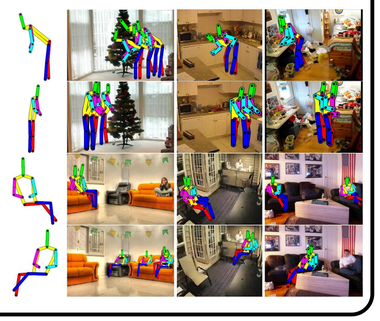
- New dataset of 146 time-lapse videos of indoor scenes.
- New statistical model describing objects by co-occurring human poses.
- Learn person-object interactions automatically from long-term observation of people.

2 – Building a robust vocabulary of poses

Use automatic pose estimation from [1] to avoid annotating poses. Need to reduce the noise:

- Specific detectors: *standing*, *sitting* and *reaching*.
- Filter detections using background subtraction and geometric filtering.
- Group remaining poses into $K^P = 32$ pose clusters by fitting a GMM (μ_k, Σ_k, π_k). The assignment vector $\boldsymbol{q}^d$ of a pose vector x^d is then:

$$\boldsymbol{q}_k^d = \frac{p(x^d|\mu_k,\Sigma_k)\pi_k}{\sum_{j=1}^{K} p\left(x^d|\mu_j,\Sigma_j\right)\pi_j}$$



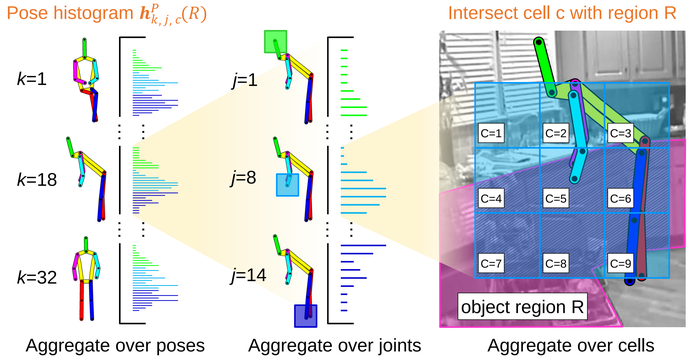
3 – Modeling person/object interactions

Describe a super-pixel R by the temporal statistics $\boldsymbol{h}^d$ of co-occurring human poses. Each detected pose $d \in D$ has bounding box $B_{j,c}^d$ for joint j and cell c, detection score s^d and weight $\boldsymbol{q}_k^d$ for each pose cluster.

The contribution of all poses to R is:

$$\boldsymbol{h}_{k,j,c}^P(R) = \sum_{d \in D} \frac{I\left(B_{j,c}^d, R\right)}{1+\exp(-3s^d)} \boldsymbol{q}_k^d$$

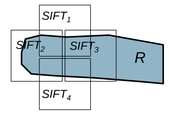
with $I(B,R) = \frac{|B \cap R|}{|B|}$.



4 – Appearance and location features

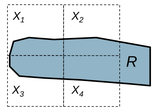
Appearance:

- Extract dense SIFT at multiple scales.
- Cluster them into $K^A = 1024$ visual words.
- Aggregate them into a bag-of-visual-words histogram $\boldsymbol{h}^A$.



Location:

- Discretize absolute image location into 10×10 cells.
- The location feature is an histogram $\boldsymbol{h}^L$ where each bin is the area of intersection of R with the corresponding absolute location cell.



5 – Learning from long-term observations

- Each super-pixel R is assigned to an object and represented by:
  $\boldsymbol{h}(R) = [\boldsymbol{h}^A(R) \cdots \boldsymbol{h}^L(R) \cdots \boldsymbol{h}^P(R)]$.
- Train linear SVM (1-vs-all, Hellinger kernel) for each object class.
- Normalize output scores by multinomial logistic regression.
- At test time: extract super-pixels and compute probability of each object class.

6 – Experiments

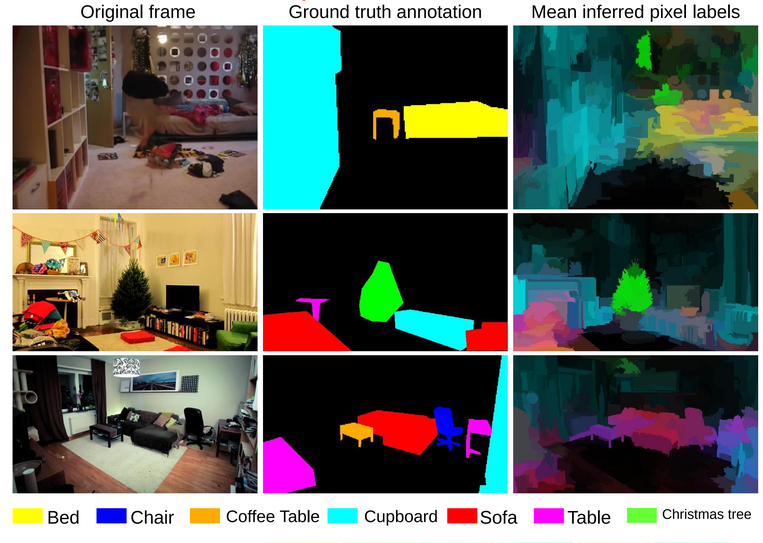
Evaluate semantic labeling of objects by pixel-wise average precision (higher is better).

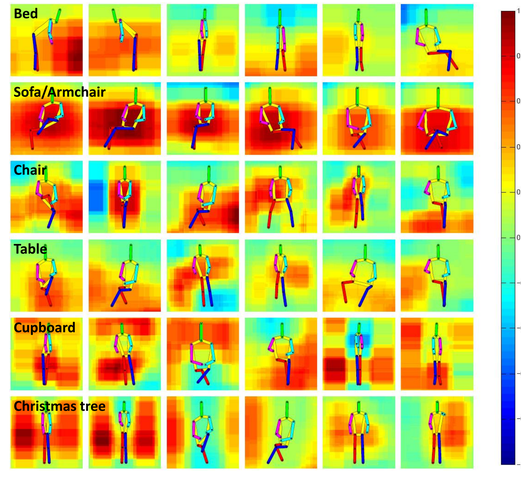
A: appearance features, **L**: location features, **P**: pose features.

	DMP [2]	Layout [3]	(A+L)	(P)	(A+P)	(A+L+P)
Wall	---	75 ± 3,9	76 ± 1,6	76 ± 1,7	**82 ± 1,2**	81 ± 1,3
Ceiling	---	47 ± 20	53 ± 8,0	52 ± 7,4	**69 ± 6,7**	**69 ± 6,6**
Floor	---	59 ± 3,1	64 ± 5,5	65 ± 3,6	**76 ± 3,2**	**76 ± 2,9**
Bed	**31 ± 20**	12 ± 7,2	14 ± 5,0	21 ± 5,8	27 ± 13	26 ± 13
Sofa/Armchair	26 ± 9,4	26 ± 10	34 ± 3,3	32 ± 6,5	**44 ± 5,4**	43 ± 5,8
Coffee Table	11 ± 5,4	11 ± 5,2	11 ± 4,4	12 ± 4,3	**17 ± 10**	**17 ± 9,6**
Chair	9,5 ± 3,9	6,3 ± 2,8	8,3 ± 2,7	5,8 ± 1,4	11 ± 5,4	**12 ± 5,9**
Table	15 ± 6,4	18 ± 3,8	17 ± 3,9	16 ± 7,1	**22 ± 6,2**	**22 ± 6,4**
Wardrobe/ Cupboard	27 ± 10	27 ± 8,2	28 ± 6,4	22 ± 1,1	**36 ± 7,4**	**36 ± 7,2**
Christmas Tree	50 ± 3,3	55 ± 12	72 ± 1,8	20 ± 6,0	76 ± 6,2	**77 ± 5,5**
Other Object	12 ± 6,4	11 ± 1,2	7,9 ± 1,9	13 ± 4,2	**16 ± 8,3**	**16 ± 8,2**
Average	23 ± 1,8	31 ± 2,0	35 ± 2,4	30 ± 1,7	**43 ± 4,4**	**43 ± 4,3**

- The proposed method outperforms the DPM [2] and Layout [3] baselines.
- The (A+P) setup significantly outperforms the (A+L) setup.
- Adding location (A+L+P) does not improve over (A+P): people already carry location information.

7 – Qualitative Results

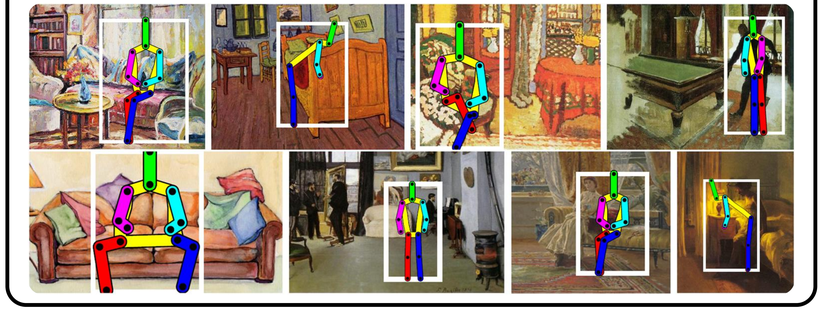




Above: ***Soft segmentations.*** Scene background with no people (left), object ground truth (middle), mean probability map for inferred objects (right).

Right: ***Spatial locations of objects relative to particular poses.*** Top 6 pose clusters with the highest sum of positive weights for selected objects (rows).

Below: ***Pose prediction.*** Select a pose cluster leading to the best agreement between the (manually provided) scene object layout and the object weights learned for each joint.



[1] Yang, Y., Ramanan, D.: Articulated pose estimation using flexible mixtures of parts. In: CVPR. (2011)

[2] Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. PAMI 32 (2010) 1627-1645

[3] Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. In: ICCV. (2009)