Scene semantics from long-term observation of people

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

Scene semantics from long-term observation of people

Vincent Delaitre[1], David F. Fouhey[2],
Ivan Laptev[1], Josef Sivic[1], Abhinav Gupta[2], and Alexei A. Efros[1,2]

[1]INRIA/École Normale Supérieure, Paris [2]Carnegie Mellon University

Abstract. Our everyday objects support various tasks and can be used by people for different purposes. While object classification is a widely studied topic in computer vision, recognition of object function, i.e., what people can do with an object and how they do it, is rarely addressed. In this paper we construct a functional object description with the aim to recognize objects by the way people interact with them. We describe scene objects (sofas, tables, chairs) by associated human poses and object appearance. Our model is learned discriminatively from automatically estimated body poses in many realistic scenes. In particular, we make use of time-lapse videos from YouTube providing a rich source of common human-object interactions and minimizing the effort of manual object annotation. We show how the models learned from human observations significantly improve object recognition and enable prediction of characteristic human poses in new scenes. Results are shown on a dataset of more than 400,000 frames obtained from 146 time-lapse videos of challenging and realistic indoor scenes.

1 Introduction

What are people expected to do with a Christmas tree just set up in a living room? Is it common to see a person sitting on a stove? Current computer vision methods provide no answers to such questions. Meanwhile, resolving these and many other questions by recognizing functional properties of objects and scenes would be highly relevant for addressing the tasks of abnormal event detection and predicting future events in image and video data.

Object functions can be derived from the known associations between object categories and human actions (the *mediated perception of function* approach [1]), for example *chair→sittable, window→openable.* Actions such as sitting, however, can be realized in many different forms which can be characteristic for some objects but not for others, as illustrated in Figure 1. Moreover, some objects may not support the common function associated with their category: for example, windows in airplanes are usually not openable. These and numerous other examples suggest that the category-level association between objects and their functions is not likely to scale well to the very rich variety of the types and forms of person-object interactions. Instead, we argue that the functional descriptions of objects should be learned directly from observations of visual data.

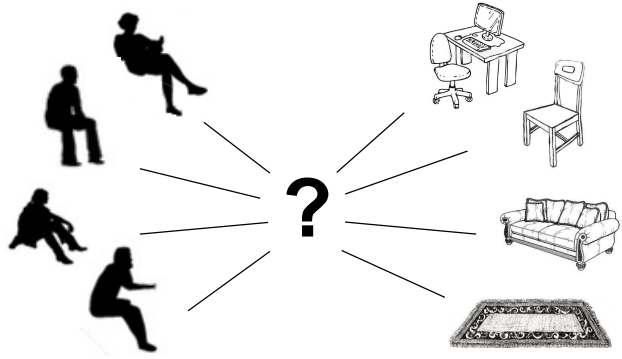

Fig. 1. *Different ways of using objects.* While all people depicted on the left are sitting, their sitting poses can be rather unambiguously associated with the objects on the right. In this paper we build on this observation and learn object descriptions in terms of characteristic body poses.

In this work we design object descriptions by learning associations between objects and spatially co-occurring human poses. To capture the rich variety of person-object interactions, we automatically detect people and estimate body poses in long-term observations of realistic indoor scenes using the state-of-the-art method of [2]. While reliable pose estimation is still a challenging problem, we circumvent the noise in pose estimation by observing *many* person interactions with the *same instances* of objects. For this purpose we use videos from hours-lasting events (parties, house cleaning) recorded with a static camera and summarized into time-lapses[1]. Static objects in time-lapses (e.g., sofas) can be readily associated with hundreds of co-occurring human poses spanning the typical interactions of people with these objects (see Figures 2-4). Equipped with this data, we construct statistical object descriptors which combine the signatures of object-specific body poses as well as the object's appearance. The model is learned discriminatively from many time-lapse videos of variety of scenes.

To summarize our contributions, we propose a new statistical model describing objects in terms of distributions of associated human poses. Notably, we do not require human poses to be annotated during training and learn the rich variety of person-object interactions automatically from long-term observations of people. Our functional object description generalizes across realistic and challenging scenes, provides significant improvements in object recognition and supports prediction of human poses in new scenes.

Background. Semantic object labeling and segmentation has been mainly considered for outdoor scenes, e.g. [3, 4]. For indoor scenes the focus has been on recovering spatial layout [5–7], possibly since many indoor objects are often better defined by their function rather than appearance.

[1] Time-lapse `http://en.wikipedia.org/wiki/Time-lapse_photography` is a common media type used to summarize recordings of long events into short video clips by temporal subsampling. We use time-lapses widely available on public video sharing web-sites such as YouTube, which are typically sampled at one frame per 1-60 seconds.

The interplay between people and objects has recently attracted significant attention. Interactions between people and semantic objects has been studied in *still images* with the focus on improving action recognition [8, 9], object localization [8, 10, 11] and discovery [12] as well as pose estimation [13, 14]. In video, constraints between human actions and objects (e.g., drinking from a coffee cup) have been investigated in restricted laboratory setups [8, 15, 16] or ego-centric scenarios [17]. In both still images and video the focus has been typically on small objects manipulated by hands (e.g., coffee cups, footballs, tennis rackets) rather than scene objects such as chairs, sofas or tables, which exhibit large intra-class variability. In addition, manual annotation of action categories [8, 16] or human poses [13] in the training data is often required and the models typically do not allow predicting poses in new scenes without people.

Functional scene descriptions have been developed for surveillance setups, e.g., [18–20], but the models are usually designed for specific scene instances and use only coarse-level observations of object/person tracks [19, 20], or approximate person segments obtained from background subtraction [18]. In contrast, our method generalizes to new challenging scenes, and uses finer grain descriptors of estimated body configuration enabling discrimination between object classes such as sofas and chairs.

Recent attempts [21, 22] have inferred functions or affordances [23] from automatically obtained noisy 3D reconstructions of indoor scenes. These methods infer affordance based on the geometry and physical properties of the space. For example, they find places where a person *can* sit by fitting a 3D human skeleton in a particular pose at a particular location in the scene. While people can sit at many places, they tend to sit in sofas more often than on tables. Moreover, they may sit on sofas in a different way than on a floor or on a chair. In this work we aim to leverage these observations and focus on *statistical affordances* by learning typical human poses associated with each object.

In a similar setup to ours, Fouhey *et al.* [7] have looked at people's actions as a cue for a coarse 3D box-like geometry of indoor scenes. Here we investigate the interplay between object function and object semantics, rather than scene geometry. In addition, in [7] the geometric person-scene relations are designed manually. In this work, we learn semantic person-object interactions from data.

2 Method overview

In this section we give a brief overview of the proposed approach. Our main goal is to learn functional object descriptions from realistic observations of person-object interactions. To simplify the learning task, we assume input videos to contain static objects with fixed locations in each frame of the video. Annotation of such objects in the whole video can be simply done by outlining object boundary in one video frame as illustrated in Figure 2. Moreover, person interactions with static objects can be automatically recorded by detecting people in the spatial proximity of annotated objects.

We start by over-segmenting input scenes into super-pixels, which will form the candidate object regions (details given in Section 5). For each object region

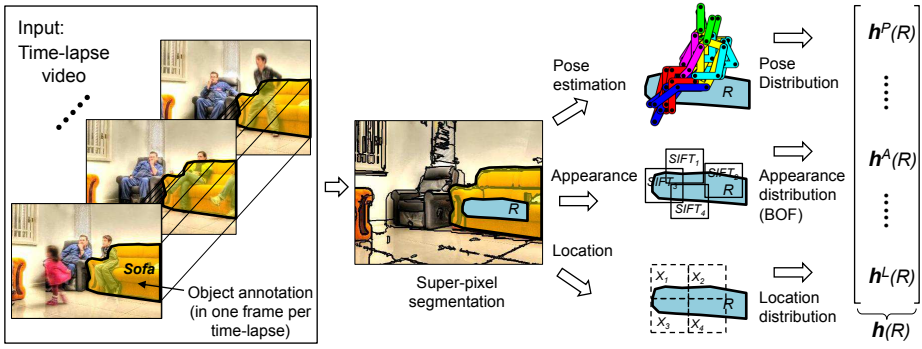

Fig. 2. *Overview of the proposed person-based object description.* Input scenes are over-segmented into super-pixels; each super-pixel (denoted R here) is described by the distribution of co-occurring human poses over time as well as by the appearance and location of the super-pixel in the image.

R we construct a descriptor vector $\boldsymbol{h}(R)$ to be used for subsequent learning and recognition. The particular novelty of our method is a new descriptor representing an object region by the temporal statistics $\boldsymbol{h}^P(R)$ of co-occurring people (Section 3). This descriptor contains a distribution of human body poses and their relative location with respect to the object region. We also represent each object region by appearance features, denoted $\boldsymbol{h}^A(R)$, and the absolute location in the frame, denoted $\boldsymbol{h}^L(R)$, as described in Section 4.

Given descriptor vectors, one for each object region, containing statistics of characteristic poses, appearance and image locations, a linear support vector machine (SVM) classifier is learnt for each object class from the labelled training data in a discriminative manner. At test time, the same functional and appearance representation is extracted from candidate object regions of the testing video. Individual candidate object regions are then classified as belonging to one of the semantic object classes.

3 Modeling long-term person-object interactions

This section presents our model of the relationship between objects and surrounding people. We start by introducing a new representation describing an object by the statistics of co-occurring human poses. We then explain the details of the extraction and quantization of human poses in time-lapses.

3.1 Describing an object by a distribution of poses

We wish to characterize objects by the typical locations and poses of surrounding people. While 3D reasoning about people and scenes [22] has some advantages, reliable estimation of scene geometry and human poses in 3D is still an open problem. Moreover, deriving rich person-object co-occurrences from a single image is difficult due to the typically limited number of people in the scene and the noise of automatic human pose estimation. To circumvent these problems, we

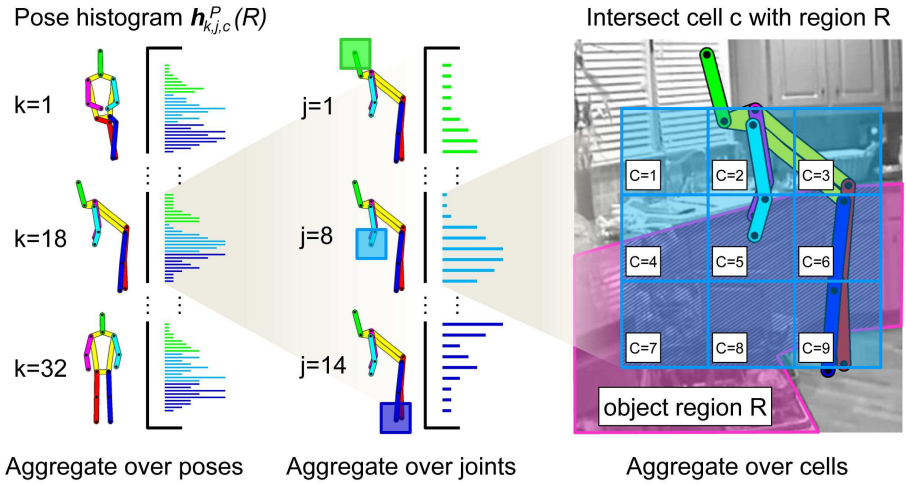

Fig. 3. *Capturing person-object interactions.* An object region R is described by a distribution (histogram) over poses k (left), joints j (middle) and cells c (right). The 3×3 grid of cells c is placed around each joint to capture the relative position of an object region R with respect to joint j. The pixel overlap between the grid cell c and the object region R weights the contribution of the j^{th} joint and the k^{th} pose cluster.

take advantage of the spatial co-occurrence of objects and people in the image plane. Moreover, we accumulate many human poses by observing scenes over an extended period of time.

In our setup we assume a static camera and consider larger objects such as sofas and tables which are less likely to change locations over time. We describe object region R in the image by the temporal statistics $\boldsymbol{h}^P$ of co-occurring human poses. Each person detection d is represented by the locations of $J(=14)$ body joints, indexed by j, and the assignment q^d_k of d's pose to a vocabulary of K^P discrete pose clusters; see Figure 3 and Sections 3.2-3.3 for details. To measure the co-occurrence of people and objects, we define a spatial grid of 9 cells c around each body joint j. We measure the overlap between the object region R and the grid cell $B^d_{j,c}$ by the normalized area of their intersection $\mathcal{I}(B_{j,c}, R) = \frac{|B_{j,c} \cap R|}{|B_{j,c}|}$. We then accumulate overlaps from all person detections $\mathcal{D}$ in a given video and compute one entry $h^P_{k,j,c}(R)$ of the histogram descriptor $\boldsymbol{h}^P(R)$ for region R as

$$h^P_{k,j,c}(R) = \sum_{d \in \mathcal{D}} \frac{\mathcal{I}(B^d_{j,c}, R)}{1 + \exp(-3s_d)}\, q^d_k, \tag{1}$$

where k, j, and c index pose clusters, body joints and grid cells, respectively. The contribution of each person detection in (1) is weighted by the detection score s_d. The values of q^d_k indicate the similarity of the person detection d with a pose cluster k. In the case of the hard assignment of d to the pose cluster $\tilde{k}$, $q^d_k = 1$ for $k = \tilde{k}$ and $q^d_k = 0$ otherwise. In our experiments we found that better results can be obtained using soft pose assignment as described in the next section.

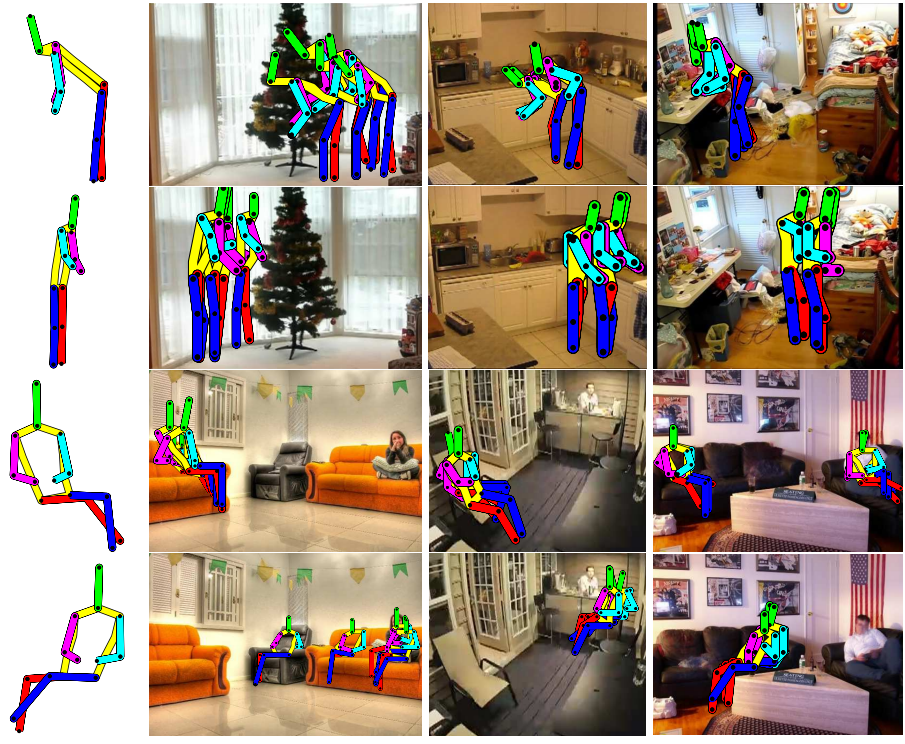

Fig. 4. *Pose cluster and detection examples.* Left: example cluster means from our pose vocabulary. Right: person detections in multiple frames of time-lapse videos assigned to the pose clusters on the left.

3.2 Building a vocabulary of poses

We represent object-specific human actions by a distribution of *quantized* human poses. To compute pose quantization, we build a vocabulary of poses from person detections in the training set by unsupervised clustering.

In order to build the pose vocabulary, we first convert each detection d in the training video into a $2J$-dimensional pose vector $\boldsymbol{x}^d$ by concatenating midpoint coordinates of all detected body joints. We center and normalize all pose vectors in the training videos and cluster them by fitting a Gaussian Mixture Model (GMM) with K^P components via expectation maximization (EM). The components are initialized by the result of a K-means clustering and during fitting we constrain the covariances to be diagonal. The resulting mean vectors $\boldsymbol{\mu}_k$, diagonal covariance matrices $\boldsymbol{\Sigma}_k$ and weights π_k for each pose cluster $k = 1, \cdots, K^P$ form our vocabulary of poses (see Figure 4). A pose vector $\boldsymbol{x}^d$ for a detection d can be described by a soft assignment to each of the $\boldsymbol{\mu}_k$ by computing the posterior probability vector $\boldsymbol{q}^d$, where

$$q_k^d = \frac{p(\boldsymbol{x}^d|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k}{\sum_{j=1}^{K^P} p(\boldsymbol{x}^d|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\pi_j}. \tag{2}$$

3.3 Person detection and pose estimation

We focus on detecting people in three body configurations common in indoor scenes: standing, sitting and reaching. We use the person detector from Yang and Ramanan [2], which was shown to perform very well at both people detection and pose estimation and train three separate models, one for each body configuration. We found that training 3 separate models improved pose estimation performance over using a single generic pose estimator (Section 7).

The three detectors are run separately on all frames of each time-lapse video in a sliding window manner at multiple scales. As all our videos have fixed viewpoint, we use background subtraction (Section 7) to remove some false positive detections. Additional false positives can be removed via geometric filtering: we use the vanishing point estimation method proposed in [24] to compute the horizon height y_h. We then assume a linear relationship $h_p(y_p) = \alpha(y_p - y_h)$ between a person's height h_p and the feet y-coordinate y_p in the image [25], and learn the scaling coefficient α via RANSAC and robust least square fitting. We discard detections for which the difference between the detected person height and the expected person height is greater than a given threshold ϵ. Finally we normalize the output of the detectors by making the mean and standard deviation of the detection scores equal to 0 and 1 on training videos, respectively. The filtering and normalization is performed separately for each detector.

To obtain the final set of detections, we perform standard non-maxima suppression on the combined outputs of the three detectors in each frame: if bounding boxes of several person detections overlap (i.e., have intersection over union bigger than 0.3), the detection with the highest normalized response is kept. This leads to a set $\mathcal{D}_i$ of confident person detections for the i^{th} video. Each detection $d \in \mathcal{D}_i$ is represented by an associated normalized score s^d and an estimated limb-configuration consisting of J bounding boxes B_j^d, $j = 1, \cdots, J$ corresponding to $J = 14$ locations of body joints.

As our time-lapse videos are sparsely sampled in time, the reasoning about temporal evolution of human poses is not straightforward. We therefore currently discard any temporal information about detected people. Nevertheless, the temporal re-occurrence of characteristic body poses for particular objects is a very powerful cue which we exploit to (i) reduce the noise in pose estimation and (ii) to span the rich variety of person-object interactions.

4 Modeling appearance and location

In addition to the distribution of poses we also model the appearance and absolute position of image regions. We build on the orderless bag-of-features representation [26] and describe the appearance of image regions by a distribution of visual words. We first densely extract SIFT descriptors [27] $f \in \mathcal{F}_k$ from image patches B^f of multiple sizes s_k for $k = 1, \cdots, S$ for all training videos and quantize them into visual words by fitting a GMM with K^A components. Each feature f is then soft-assigned to this vocabulary in the same manner as described in Eq. (2). This results in an assignment vector $\boldsymbol{q}^f$ for each feature.

The K^A-dimensional appearance histogram $\boldsymbol{h}^A(R)$ for region R is computed as a weighted sum of assignment vectors $\boldsymbol{q}^f$

$$\boldsymbol{h}^A(R) = \sum_{k=1}^{S} \sum_{f \in \mathcal{F}_k} s_k^2 \mathcal{I}(B^f, R)\, \boldsymbol{q}^f, \tag{3}$$

where $s_k^2\mathcal{I}(B^f, R)$ is the number of pixels belonging to both object region R and feature patch B^f.

Similar to [28], we also represent the absolute position of regions R within the video frame. This is achieved by spatially discretizing the video into a grid of $m \times n$ cells, resulting in a $(m \times n)$-dimensional histogram $\boldsymbol{h}^L(R)$ for each region R. Here the i^{th} bin of $\boldsymbol{h}^L(R)$ is simply the proportion of pixels of the i^{th} cell of the grid falling into R.

5 Learning from long-term observations

We now detail how we obtain candidate object regions from multiple super-pixel segmentations and learn the model of person-object interactions. We then show how to recognize objects in testing videos and predict likely poses in new scenes.

Obtaining candidate object regions. As described in previous sections, we represent objects by accumulating statistics of human poses, image appearance and location at object regions R. Candidate object regions are obtained by over-segmenting video frames into super-pixels using the method and on-line implementation of [29]. As individual video frames may contain many people occluding the objects in the scene, we represent each video using a single "background frame" containing (almost) no people (Section 7). Rather than relying on a single segmentation, we follow [28] and compute multiple overlapping segmentations by varying the parameters of the segmentation algorithm.

Learning object model. We train a classifier for each object class in a one-versus-all manner. The training data for each classifier is obtained by collecting all (potentially overlapping) super-pixels, R_i for $i = 1, \cdots, N$, from all training videos. For each region, we extract their corresponding pose, appearance and location histograms as described in Sections 3 and 4. The histograms are separately L_1-normalized and concatenated into a single K-dimensional feature vector $\boldsymbol{x}_i = [\tilde{\boldsymbol{h}}^P(R_i), \tilde{\boldsymbol{h}}^A(R_i), \tilde{\boldsymbol{h}}^L(R_i)]$, where $\tilde{\boldsymbol{h}}$ denotes L_1-normalized histogram $\boldsymbol{h}$. An object label y_i is then assigned to each super-pixel based on the surface overlap with the provided ground truth object segmentation in the training videos. Using the surface overlap threshold of 34%, each super-pixel can be assigned up to two ground truth object labels. Finally we train a binary support vector machine (SVM) classifier with the Hellinger kernel for each object class using the labelled super-pixels as training data. The Hellinger kernel is efficiently implemented using the explicit feature map $\Phi(\boldsymbol{x}_i) = \sqrt{\boldsymbol{x}_i / L_1(\boldsymbol{x}_i)}$ and a linear classifier. Finally, the outputs of individual SVM classifiers are calibrated with respect to each other by fitting a multinomial regression model from the classifiers output to the super-pixel labels [30]. The output of the learning stage is

a K-dimensional weight vector $\boldsymbol{w}_y$ of the (calibrated) linear classifier for each object class y.

At test time, multiple super-pixel segmentations are extracted from the background frame of the test video and the individual classifiers are applied to each super-pixel. This leads to a confidence measure for each label and super-pixel. The confidence of a single image pixel is then the mean of the confidences of all the super-pixels it belongs to.

Inferring probable pose. Here we wish to predict the most likely pose within a manually provided bounding box in an image, given an object layout (segmentation) of the scene. This is achieved by choosing the pose cluster, for which the sum of learnt object weights for all joints most agree with the given per-pixel object labels in the image. More formally, denoting $w_y(k, j, c)$ the weight learnt for label y, pose cluster k, joint j and grid cell c, we select the pose cluster $\hat{k}$ that maximizes the sum of per-pixel weights under each joint grid cell $B^k_{j,c}$

$$\hat{k} = \arg\max_k \sum_{j=1}^{J} \sum_{c=1}^{9} \sum_{\text{pixels } i \in B^k_{j,c}} w_{y_i}(k, j, c), \tag{4}$$

where y_i is the label for pixel i.

6 Time-lapse dataset

We extend the dataset of [7] to 146 time-lapse videos containing a total of around 400,000 frames. Each video sequence shows human actors interacting with an indoor scene over a period of time ranging from a few minutes to several hours. The captured events include parties, working in an office, cooking or room-cleaning. The videos were downloaded from YouTube by placing queries such as "time-lapse party". Search results were manually verified to contain only videos captured with a stationary camera and showing an indoor scene. All videos are sparsely sampled in time with limited temporal continuity between consecutive frames. The dataset represents a challenging uncontrolled setup, where people perform natural non-staged interactions with objects in a variety of real indoor scenes.

We manually annotated each video with ground truth segmentation masks of eight frequently occurring semantic object classes: *'Bed'*, *'Sofa/Armchair'*, *'Coffee Table'*, *'Chair'*, *'Table'*, *'Wardrobe/Cupboard'*, *'Christmas tree'* and *'Other'*. Similar to [24], the *'Other'* class contains various foreground room clutter such as clothes on the floor, or objects (e.g., lamps, bottles, or dishes) on tables. In addition to objects we also annotated three room background classes: *'Wall'*, *'Ceiling'* and *'Floor'*. As the camera and majority of the objects are static, we can collect hundreds or even thousands of realistic person-object interactions throughout the whole time-lapse sequence by providing a single object annotation per video. The dataset is divided into 5 splits of around 30 videos with approximately the same proportion of labels for different objects. The dataset including the annotations is available at `http://www.di.ens.fr/willow/research/scenesemantics/`.

7 Experiments

In this section we give the implementation details and then show results for (i) pose estimation (ii) semantic labeling of objects in time-lapse videos and (iii) predicting likely poses for new scenes.

Implementation details. The foreground/background segmentation in each video frame is estimated using a pixel-wise adaptive mixture of Gaussian with 5 components [31] (with $\alpha = 0.01$ and $T = 0.2$). We also compute a single "background image" for each video that contains no people by taking the median of background segments across all video frames. Person detections and human pose estimates in each frame are obtained using the method and code of [2]. Detections in the background segments and with confidence smaller than -1.1 are removed. The threshold ϵ for the ground-plane based geometric filter [25] is set to 30%. Super-pixels for each video are generated using the code of [29] with parameters $\sigma \in \{0.2, 0.3\}$, $k = 80$ and $min = 600$. SIFT features are extracted from patches of size $s \in \{8, 16, 32, 64\}$ pixels, with 50% spatial overlap. To train the proposed model, we use 3 splits of the dataset (see section 6) to cross-validate the C parameter of the SVM and use the 4^{th} split to calibrate the outputs of the individual classifiers. The resulting model is tested on the 5^{th} split. This is repeated five times for the different test splits to obtain the mean and standard deviation of the classification performance.

Pose estimation. To evaluate person detection and pose estimation performance we have annotated poses of at least ten (randomly chosen) person occurrences in each video, resulting in 1606 pose annotations. Person (bounding box) detection performance is measured using the standard average precision (AP) and pose estimation performance is measured by the Percentage of Correct Parts (PCP) score among the detected people as proposed in [32]. We first compare our individually trained pose estimators for each action (see section 3.3) with a single model trained on images from all 3 action classes. Both have a similar recall of around 52% but the individually trained models achieve an average PCP of 50% compared to 47% for the single model. We then evaluate the effect of the background subtraction and geometric filtering for person detection. The individually trained models achieve an AP of 33%, which is significantly improved by background subtraction (51%) and geometric filtering (56%).

Semantic labeling of objects. Semantic labeling performance is measured by pixel-wise precision-recall curve and average precision (AP) for each object. Table (1) shows the average precision for different object and room background classes for different feature combinations of our method. Performance is compared to two baselines: the method of [24], trained on our data with semantic object annotations, and the deformable part model (DPM) of [33] trained over manually defined bounding boxes for each class. At test time, the DPM bounding boxes are converted to segmentation masks by assigning to each testing pixel the maximum score of any overlapping detection. Note that combining the proposed pose features with appearance (A+P) results in a significant improvement

	DPM [33]	[24]	(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**

Table 1. Average precision (AP) for baselines of Felzenszwalb*et al.* [33] and Hedau*et al.* [24] compared to four different settings of our method: appearance and location features only (A+L), person features only (P), appearance and person features (A+P), appearance, location and person features combined (A+L+P).

in overall performance, but further adding location features (A+L+P) brings little additional benefit, which suggests that spatial information in the scene is largely captured by the spatial relation to the human pose. The proposed method (A+L+P) also significantly outperforms both baselines. Example classification results for the proposed method are shown in Figure 5. Finally, learnt weights for different objects are visualized in Figure 6.

We have also evaluated our model on functional surface estimation. For training and testing, we have provided ground truth functional surface masks for the dataset of [7]. Our model achieves AP of 76%, 25% and 44% for *'Walkable'*, *'Sittable'* and *'Reachable'* surfaces, respectively, averaging a gain of 13% compared to [7], which could be attributed to the discriminative nature of our model.

Predicting poses in new scenes. Figure 7 shows qualitative results of predicting likely human poses in new scenes. Given a person bounding box and the manually labelled object regions, the most likely pose is predicted using Eq. (4). As can be seen, the automatically generated poses are consistent with object classes as well as with the scene geometry despite no explicit 3D reasoning is included in our model.

8 Discussion

We have proposed a statistical descriptor of person-object interactions and have demonstrated its benefits for recognizing objects and predicting human body poses in new scenes. Notably, our method requires very little annotation and relies on long-term observations of people in time-lapse videos. Given the mutual dependence of objects and human poses, the current method can be further extended to perform joint pose estimation and object recognition.

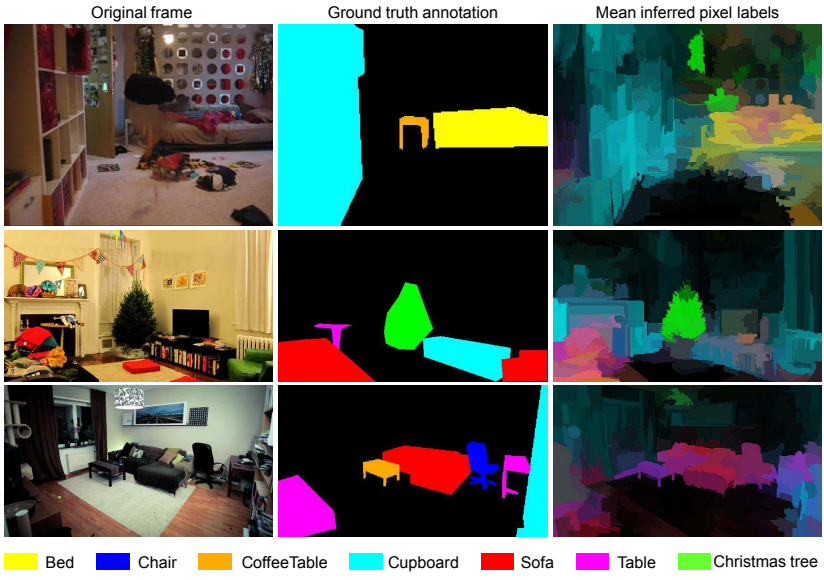

Fig. 5. *Object soft segmentation.* Scene background with no people (left). Object ground truth (middle). Mean probability map for inferred objects (right).

Acknowledgments: This work was supported by a NSF Graduate Research Fellowship to DF, by ONR-MURI N000141010934, ONR Grant N000141010766, Quaero program funded by OSEO, MSR-INRIA, EIT-ICT and ERC grant Videoworld.

References

1. Palmer, S.E.: Vision science: photons to phenomenology. MIT Press, Cambridge, Mass. (1999)
2. Yang, Y., Ramanan, D.: Articulated pose estimation using flexible mixtures of parts. In: CVPR. (2011)
3. Kohli, P., Ladicky, L., Torr, P.: Robust higher order potentials for enforcing label consistency. IJCV **82** (2009) 302–324
4. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. ECCV (2006)
5. Hedau, V., Hoiem, D., Forsyth, D.: Thinking inside the box: Using appearance models and context based on room geometry. In: ECCV. (2010)
6. Lee, D., Hebert, M., Kanade, T.: Geometric reasoning for single image structure recovery. In: ICCV. (2009)
7. Fouhey, D., Delaitre, V., Gupta, A., Efros, A., Laptev, I., Sivic, J.: People watching: Human actions as a cue for single view geometry. In: ECCV. (2012)
8. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: Using spatial and functional compatibility for recognition. PAMI (2009)
9. Delaitre, V., Sivic, J., Laptev, I.: Learning person-object interactions for action recognition in still images. In: NIPS. (2011)

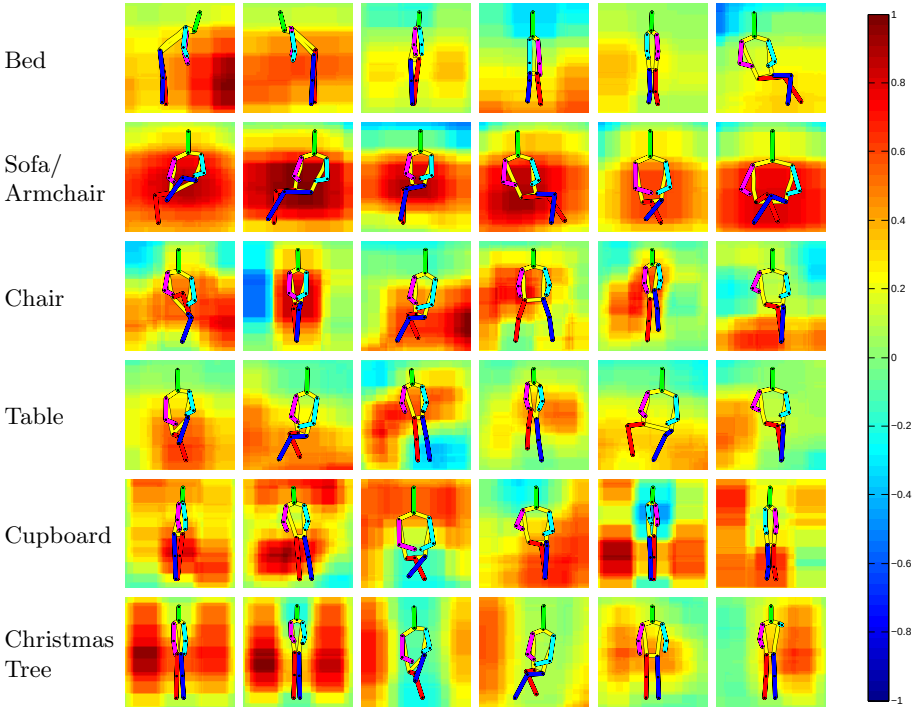

Fig. 6. *Spatial locations of objects relative to particular poses.* The top 6 pose clusters with the highest sum of positive weights are shown for selected objects (rows). Color indicates the spatial weights for the position of a given object relative to the particular body pose summed over all 9 grid cells for all joints. The color map is shown on the right. Note how, for example, Sofa/Armchair is likely to be located behind sitting people (2nd row) and table in the vicinity of sitting and standing people (4th row). The top scoring sitting poses for Sofa/Armchair are also quite different (more relaxed) than the top scoring sitting poses for Chair.

10. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for static human-object interactions. In: SMiCV, CVPR. (2010)
11. Stark, M., Lies, P., Zillich, M., Wyatt, J., Schiele, B.: Functional object class detection based on learned affordance cues. In: ICVS. (2008)
12. Prest, A., Schmid, C., Ferrari, V.: Weakly supervised learning of interactions between humans and objects. PAMI (2011)
13. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR. (2010)
14. Yao, B., Khosla, A., Fei-Fei, L.: Classifying actions and measuring action similarity by modeling the mutual context of objects and human poses. In: Proc. ICML. (2011)
15. Gall, J., Fossati, A., van Gool, L.: Functional categorization of objects using real-time markerless motion capture. In: CVPR. (2011)
16. Kjellstrom, H., Romero, J., Martinez, D., Kragic, D.: Simultaneous visual recognition of manipulation actions and manipulated objects. In: ECCV. (2008)
17. Fathi, A., Ren, X., Rehg, J.: Learning to recognize objects in egocentric activities. In: CVPR. (2011)

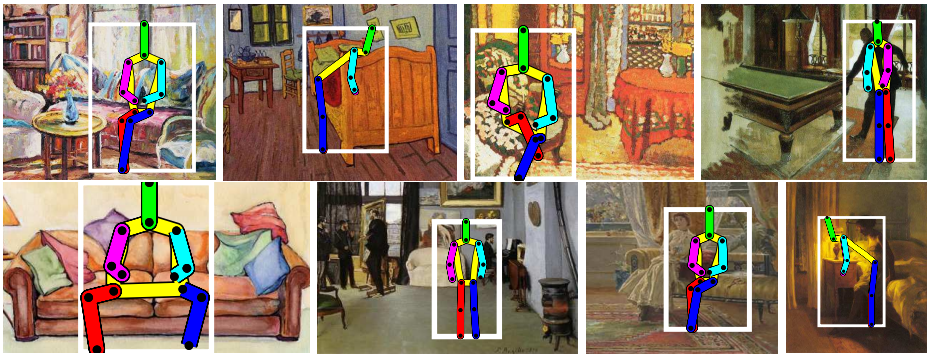

Fig. 7. *Plausible poses prediction.* The proposed model supports automatic prediction of plausible human poses in new scenes. This is achieved by selecting a pose cluster leading to the best agreement between the (manually provided) scene object layout and the object weights learned for each joint.

18. Peursum, P., West, G., Venkatesh, S.: Combining image regions and human activity for indirect object recognition in indoor wide-angle views. In: ICCV. (2005)
19. Turek, M., Hoogs, A., Collins, R.: Unsupervised learning of functional categories in video scenes. In: ECCV. (2010)
20. Wang, X., Tieu, K., Grimson, E.: Learning semantic scene models by trajectory analysis. In: ECCV. (2006)
21. Grabner, H., Gall, J., Van Gool, L.: What makes a chair a chair? In: CVPR. (2011)
22. Gupta, A., Satkin, S., Efros, A., Hebert, M.: From 3d scene geometry to human workspace. In: CVPR. (2011)
23. Gibson, J.: The ecological approach to visual perception. Boston: Houghton Mifflin (1979)
24. Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. In: ICCV. (2009)
25. Rodriguez, M., Laptev, I., Sivic, J., Audibert, J.Y.: Density-aware person detection and tracking in crowds. In: ICCV. (2011)
26. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: WS-SLCV, ECCV. (2004)
27. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
28. Hoiem, D., Efros, A., Hebert, M.: Geometric context from a single image. In: ICCV. (2005)
29. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. IJCV **59** (2004) 167–181
30. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning. Springer (2003)
31. Staufer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: CVPR. (1998)
32. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. CVPR (2008)
33. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. PAMI **32** (2010) 1627–1645