Predicting Functional Regions of Objects

Chaitanya Desai
University of California, Irvine
desaic@uci.edu

Deva Ramanan
University of California, Irvine
dramanan@ics.uci.edu

Abstract

We revisit the notion of object affordances, an idea that speaks to an object's functional properties more than its class label. We study the problem of spatially localizing affordances in the form of 2D segmentation masks annotated with discrete affordance labels. For example, we use affordance masks to denote on what surfaces a person sits, grabs, and looks at when interacting with a variety of everyday objects (such as chairs, bikes, and TVs). We introduce such a functionally-annotated dataset derived from the PASCAL VOC benchmark and empirically evaluate several approaches for predicting such functionally-relevant object regions. We compare "blind" approaches that ignore image data, bottom-up approaches that reason about local surface layout, and top-down approaches that reason about structural constraints between surfaces/regions of objects. We show that the difficulty of functional region prediction varies considerably across objects, and that in general, top-down functional object models do well, though there is much room for improvement.

1. Introduction

> "If you know what can be done with a ... object, what it can be used for, you can call it whatever you please"
> J. J Gibson [14]

Gibson eloquently argues that predicting functional "affordance" is more important than predicting object category labels. However, the vast majority of work on object recognition focuses on the task of predicting bounding boxes and category labels - see, for example, the PASCAL VOC benchmark [7]. As an example, consider the objects in Fig. 1; though it is unclear if they should be labeled as a "chair", most people would know how to sit on them. If a humanoid robot were to be confronted with these objects, it would not suffice to simply name them or estimate their bounding boxes; rather the crucial bit is knowing where the robot should rest its bum and back.

Figure 1. Objects that can potentially be used as chairs by humans. Humanoid robots, when faced with such objects would need precise localization of the regions that they can sit on (yellow) and rest their back against (blue). We benchmark a wide variety of algorithms for producing such outputs, including blind baselines that ignore image data, bottom-up models of surface geometry, and top-down models that reflect object-specific structural constraints.

We argue such precise modes of interaction exist for virtually any object category. When interacting with a bottle, we must estimate where to grab it with our hands and where to place our mouths. When interacting with a computer, we must estimate where to look, since a rear-facing monitor affords little use to an observer. The central thesis of this work is that *functional regions* are an important type of output that recognition systems should produce, alongside classic outputs as categorical labels and attribute values. We define a generic set of affordance labels based on body parts that touch an object during typical interactions (e.g., when using a bike, one places feet on pedals and hands on handlebars). Additionally, we define "looking at" as an important interaction that does not involve touching. We show examples of functional regions for everyday objects in Fig. 2.

Functional prediction dataset: Formally, we define the task of function region prediction as the prediction of segmentation masks with discrete affordance labels. We define a candidate mask and label to be correct if it overlaps the correspondingly-labeled ground-truth segmentation mask by a sufficient amount. For simplicity, we consider the case when an object bounding box is known at test-time, similar to the formulation of attribute prediction [9].

Benchmark evaluation: We compare several baseline approaches to functional region prediction. We first consider "blind" baselines that do not look at any image data, and just use the bounding box to predict functional region masks. We show that such baselines do well for certain objects with little variability in 3D structure or pose. For example, bottles tend to mostly be upright, in which case one

978-0-7695-4990-3/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPRW.2013.141

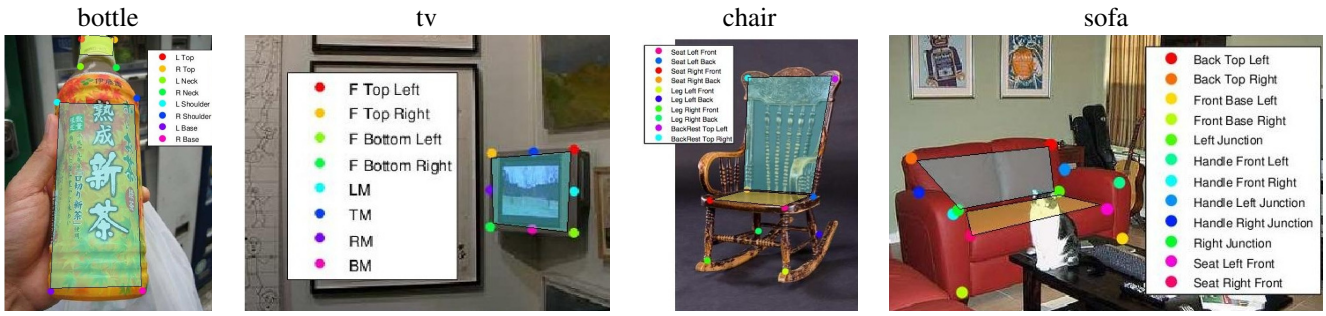

Figure 2. We show functional regions for a variety of everyday objects, visualized as translucent polygons. We derive these region labels by first annotating objects with functional landmarks that define polygon corners.

can simply "grasp" the bottom of an instance's bounding box and "put their mouth" on the top. For other objects such as chairs, functional region prediction is much more nuanced. We introduce novel but simple approaches based on bottom-up geometric surface layout estimation, as well as object-specific top-down models. Bottom-up geometric models are effective for coarse-scale layout estimation [17]. However, top-down models can take advantage of object-specific structural constraints; e.g., for a chair, the back rest is above and perpendicular to the seat. We show that such high-level constraints are important for good performance.

2. Related Work

Object affordances: J.J. Gibson coined the term affordances to describe object function [14], though such notions date back at least to the Gestalt school of psychology [18]. Early computer vision research relating to object function include [23, 24, 26]. [23] describe methods for estimating the function of known objects by reasoning about their constituent parts and relations; for example, a hammer can be described by a handle connected to an end effector that strikes targets. We explore top-down models that similarly connect object shape to function, but our models are learned from data rather than hand-coded. Along these lines, [15] learn a "sittable" affordance detector of a chair by fitting 3D models of a sitting human skeleton to 3D models of chairs. [27] describe a method for computing planar surface approximations of everyday objects using view-based deformable models. While they evaluate landmark prediction, we focus on affordance region prediction.

Scene affordances: More recently, scene affordances in indoor settings have been explored in [13, 16, 19]. [13, 16] restrict the scene to a box-shaped room and estimate its 3D layout [20]. [16] use the estimated layout and occupancy model to fit cuboidal models of objects, which in turn define a functional human workspace. Cuboidal approximations of chairs and sofas are likely too coarse to resolve our desired affordance labels (that specify where to rest one's back and bum). [13] estimate functional regions by observing human actors interacting with objects in the scene - one can infer that an object is "sittable" because multiple people have sat on it. Our formulation differs in that we focus on estimating affordance labels directly from a static image. Presumably such reasoning is required in order estimate functional affordances when presented with a *novel* scene of objects.

Spatially-defined attributes: Attributes are another framework for reporting "interesting" properties of an object. Much work formulates attribute prediction as a discrete multilabel problem [9]. Often attributes are not tied to particular spatial regions, though [2, 6, 25] consider spatially-localized attributes, such as the type of nose in an image of a face. Our work can viewed as similar in spirit, in that we spatially localize *functionally*-important attributes of an object.

Supervised part models: Our top-down models are based on exemplar-based templates [21] and 2D pictorial structures [11]. We show that nearest-neighbor classification, followed by *functional* label transfer, is a surprisingly effective approach. We also explore deformable part models (DPMs) [12] and variants [28, 5] that are tuned to report detailed spatial reports of objects rather than just bounding boxes. This is in contrast to other supervised part models [1, 3] that ignore part localizations when reporting final output at test time. Our functional perspective also addresses one classic difficulty with supervision; it is often not clear what are the right parts - for example, what are the natural parts of a sofa? We argue that one should define parts that are necessary to understand how one *interacts* with an object. From this perspective, an armrest is a good part because it is relevant for functional region prediction. In some cases, functionally-relevant parts may look quite different from classically-defined parts; for example, part-based car detectors typically do not model the door handle, but such a landmark is extremely relevant from our functional view.

3. Blind baselines

Recall our problem formulation; we are given training images of an object with functional region masks, and we wish to predict the same region masks on test images (with bounding box annotations). In this section, we describe two

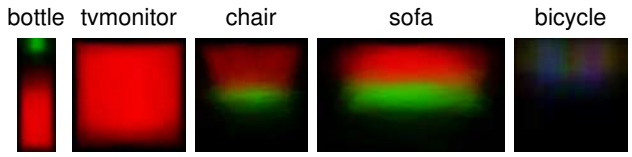

Figure 3. Blind prediction: we show pixel-level spatial priors of functional regions of `bottles` (grasp and mouth-placement), `tvs` (screen), `chair`/`sofas` (back rest and seat), and `bikes` (left handlebar, right handlebar and seat). The strong spatial priors for `bottle` and `monitor` suggest that a "blind" algorithms based exclusively on the prior may do well. Bicycles vary considerably in 3D pose, so the pixel-wise prior looks to be quite weak.

simple "blind" baseline models that do not process any pixel data. Surprisingly, we show that for some objects, such blind baselines work well.

Linear Regression: Our first blind baseline makes use of a polygonal representation of the region masks. We represent each region mask using a four-sided polygon, parameterized with 4 (x, y) corner points. We can then treat the problem of region prediction as a multivariate regression problem, where one is predicting 8 values for each affordance region of interest, using simply the bounding box coordinates as input. Let us write $(x_1^n, y_1^n, x_2^n, y_2^n)$ for the bounding box coordinates of the n^{th} training example, written in normalized image coordinates such that each training image has unit height and width. We define a linear regression problem that predicts the i^{th} corner point on the n^{th} training instance, written as p_i^n, given the height and width of the n^{th} bounding box:

$$\underset{W_i}{\text{argmin}} \sum_n ||W_i \begin{bmatrix} x_2^n - x_1^n \\ y_2^n - y_1^n \end{bmatrix} - p_i^n||^2 \qquad (1)$$
$$\text{where} \quad W_i \in \mathcal{R}^{2\times 2}, p_i^n \in \mathcal{R}^{2\times 1}$$

The above model predicts the i^{th} corner point using the aspect ratio and scale of the given bounding box. One can solve for each row of W_i independently using linear regression. We found this regression model to outperform one based on the four bounding box coordinates (probably because the above model has fewer parameters and so is less likely to overfit). At the other extreme, one might try to learn a scale-invariant predictor (or moreover, a anisotropic scale-invariant predictor invariant to aspect changes) by normalizing each bounding box to have unit height and weight, but we saw worse performance. This suggests that both the scale and aspect ratio of the bounding box are useful cues for object function. This makes sense; the aspect ratio of a couch gives clues as to its 3D pose, which in turn impacts the prediction of where we should sit and rest our backs.

Pixel-level prior: Another reasonable blind baseline maybe a pixel-level prior for the affordance label, based strictly on the location of the pixel inside the bounding box. For ease of exposition, let us first define our pixel-level prior to be aspect and scale invariant (unlike our linear regression model above). To do so, we write $x_i \in \{0, \ldots, K\}$ for the affordance label of pixel at location i, where i are coordinates within a rescaled bounding box with unit height and width. We define 0 to be the background label. Specifically, we write the joint distribution over all pixels $x = \{x_i\}$ as:

$$P(x) = \prod_i P(x_i) \quad \text{such that} \quad P(x_i = k) = \beta_{ik} \qquad (2)$$

where β_{ik} is the prior that pixel i taken on value k (such that $\sum_k \beta_{ik} = 1$). This can be readily learned by counting the frequency of labels in training data. See Fig 3 for a visualization of such learned priors.

Aspect-specific prior: We also learned aspect-specific pixel-level priors (by clustering training data based on aspect, and learning a separate model for each cluster), but saw little improvement. Our pixel-level model is much higher dimensional than our linear regression model, and so we posit that it requires additional training data to avoid overfitting.

Inference with spatial-coherence: The above prior model assumes the label of each pixel is independent of all other pixel labels, given the bounding box coordinate frame. One might also define spatial coherence model that biases neighboring pixels to have similar labels (say, with a pairwise Markov random field). Inference and learning with such a model can be difficult if pairwise potentials are not submodular. Instead, we enforce spatial coherence by requiring all pixels in a superpixel S (produced with an over-segmentation of the image [10]) to share the same affordance label.

$$\text{Label}(S) = \underset{k}{\text{argmax}} \prod_{i \in S} P(x_i = k) \qquad (3)$$

Technically speaking, the above spatial coherence model is no longer "blind" since image evidence is used to group pixels in the over-segmentation, but the above model still fails to use image evidence to decide the label of each superpixel.

4. Bottom-up model of surface-layout

We now describe bottom-up models that produce function region predictions by processing image data in a bottom-up manner.

Surface layout prediction: The functional regions of many objects can be described as planar surfaces. In this section, we show how one can use geometric surface layout models (e.g.[17]), to generate functional region predictions of such planar regions. We apply these models to a subset of our objects, including `chairs`, `sofas` and `tv monitors`. The output of [17], tuned for indoor surfaces, produces a distribution over 7 geometric classes at

each pixel: *support*, *sky*, *vertical left*, *vertical center*, *vertical right*, *vertical porous*, *vertical solid*. These models are trained using a variety of features including color, spatial position, texture features, and fitted line features. Let g_i be the geometric surface label of each pixel.

$$P(g) = \prod_i p(g_i) \qquad \text{such that} \qquad P(g_i = l) = \gamma_{il} \quad (4)$$

where γ_{il} is the prior that pixel i taken on geometric label l. To report functional labels, we define mapping function

$$P(x_i = k) \propto P(g_i = \text{map}(k)) \quad (5)$$

where map(k) maps an affordance label to a geometric class label.We take the surface labels and map them to functional labels as follows: For `chair` and `sofa`, *support* pixels are mapped to seats and *vertical* labels are mapped to backrests. Because some geometric classes contain multiple subclasses (such as *vertical*), we solve for the best mapping of subclasses to the desired functional label, so as to maximize benchmark performance (which happens when the predicted regions tend to overlap the ground-truth). Fig. 4 shows some examples of surface labels and their mapped affordance labels. As in (3), we use superpixels to enforce spatial coherence in a computationally efficient manner.

Prior+Surface: We train a model that combines the prior spatial distribution of (2) with the geometric cues of (4) using a conditional model. To do so, let us define a feature vector at each pixel f_i consisting of the probabilities returned by each model:

$$f_i = \begin{bmatrix} \beta_{i1} \dots \beta_{iK} & \gamma_{i1} \dots \gamma_{i7} \end{bmatrix}^T \quad (6)$$

These features are fed into a K-way "soft-max" classifier, trained using maximum-likelihood:

$$P(x_i = k) = \frac{e^{(\theta_k \cdot f_i)}}{\sum_l e^{(\theta_l \cdot f_i)}} \quad (7)$$

Finally, we enforce spatial coherence by forcing all pixels in a superpixel to have the same label, as in (3).

5. Top-down appearance models

We now explore top-down models that explicitly score appearance (and hence are not "blind").

Nearest-neighbor prediction (NN): We first begin with a simple nearest-neighbor approach; given a test object, we find the closest training example (in terms of appearance) and simply transfer the functional region label:

$$\text{Label}(I) = \text{Label}(i^*) \quad \text{where} \quad i^* = \underset{i}{\text{argmin}} \, ||\phi(I) - \phi(X_i)||^2 \quad (8)$$

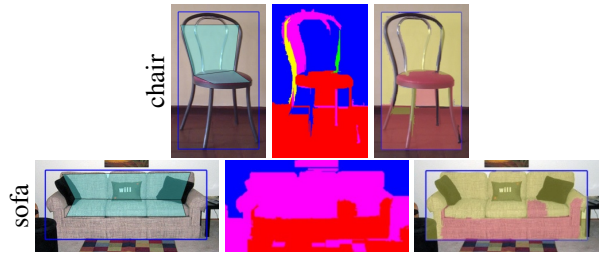

Figure 4. Bottom-up prediction: We show functional labels predicted by our **Surface** algorithm on test images. The left image shows the ground-truth functional labeling of an image (with a translucent blue mask), the middle image shows surface labels as computed by [17], and the last image shows functional labels produced by the computed mapping, where yellow corresponds to "backrest" and red correponds to "seat". Surface labels correspond to support, vertical left, vertical center, vertical right, vertical porous, vertical solid, and sky (white). Only pixels inside the object bounding box are considered for quantitative evaluation. Such bottom-up methods do well when functional regions have clear boundaries, but sometimes struggle to differentiate subtle geometry (such as the sofa seat versus backrest).

where I is the test instance, X_i is the i^{th} training instance, and Label(i) is its functional-region label. To model appearance, we resize training and test data to a canonical width and height and compute a histogram of oriented gradients (HOG) descriptor ϕ. We map the region label to the width and height of the test instance I. Similar ideas that consider nearest neighbors in terms of descriptor distances have also been used in detecting salient regions in images [22]. One would expect such a non-parametric model to need large amounts of training data to do well, but we find it is competitive even on moderately-sized training sets.

Latent DPM (LDPM): Deformable part models appear to represent the current state-of-the-art in recognition. We train a model using [12], which learns latent parts given object bounding-boxes. This model produces aspect mixture labels and part localizations associated with each detection. We train a post-processing linear regressor analogous to (1) that predicts corners of functional regions given the output of LDPM detector. Interestingly, we found the mixture component to be a more reliable cue than the part locations. As such we train a separate linear regressor using the same feature set as (1), for each mixture. We do this by re-running the learned model on the training images, and recording the input and target features for learning a regressor.

Functional DPM: We posit that functional landmarks may provide additional supervision that can be used to learn more semantically-meaningful parts. For example, it may be much easier to predict the location of a handlebar if one explicitly trains a handlebar part. We define parts at each functional landmarks that define corners of function regions (as shown in Fig. 5). Some objects have clear semantic landmarks; a `bicycle` has the left/right handle, 2 wheels, a seat, etc. However, objects such as a `chair`,

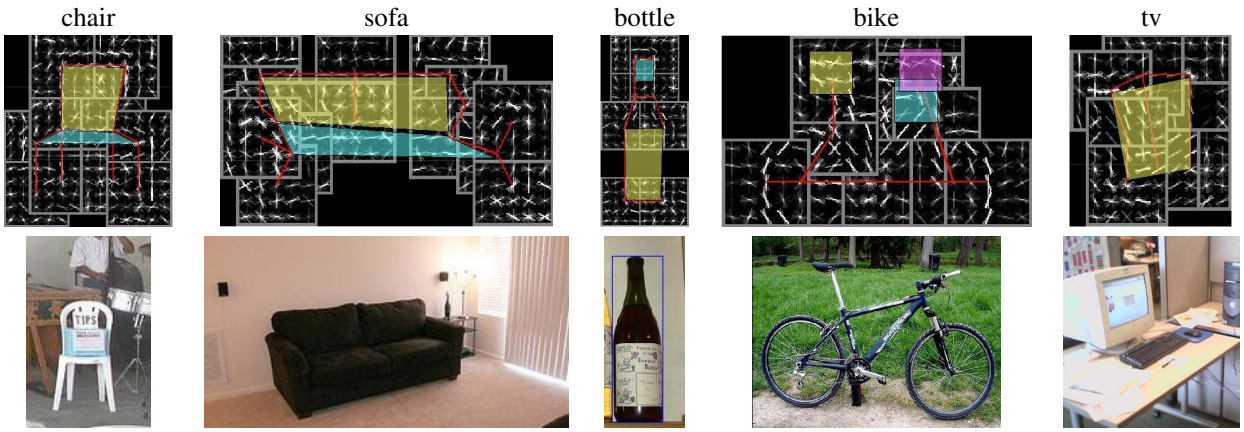

Figure 5. Top-down prediction: we visualize FunctionalDPMs and their associated functional regions. FunctionalDPMs are part models learned from functionally-supervised parts. We show gradient-based part templates along with example images that triggered a detection.

`sofa`, `bottle`, `tv monitor` are difficult to break up into parts. Instead, we define functional landmarks at corners of functional regions. As we describe in Sec. 6, we assume functional regions are represented as polygons, and simply define a landmark at each corner point (shown in Fig. 2). In fact, such keypoint annotations for a variety of PASCAL categories have been made publicly available by [3]. Fig. 2 shows examples of keypoints provided by [3] for four different objects along with their functional regions (represented as translucent segmentation masks). For each object category, we now have a training set with annotated functional landmarks.

Latent parts need not be semantically-meaningful, and tend to be coherent in appearance due to their construction. Functional parts can vary greatly in appearance due to changes in object viewpoint and structure. To model these variations, we use local part mixtures corresponding to clusters of landmark configurations, as in [5, 28]. We briefly review the formulation from [5, 28] here. Each part i is parameterized by its position $p^i = (x, y)$ and mixture type $t_i \in \{1, 2, \dots M\}$. Given an image I, we score a particular arrangement of K parts $p = \{p^i\}_{i=1:K}$ and mixture types $t = \{t^i\}_{i=1:K}$ with:

$$S(I, p, t) = \sum_{i=1}^{K} \alpha^i_{t^i} \cdot \phi(I, p^i) + \sum_{i,j \in E} \beta^{ij}_{t^i, t^j} \cdot \psi(p^i - p^j) \quad (9)$$

The first term computes the score of placing template $\alpha^i_{t^i}$, tuned for mixture t^i for part i, at location p^i. We write $\phi(I, p^i)$ for a HOG feature vector [4] extracted from pixel location p^i. We write $\psi(p^i - p^j) = \begin{bmatrix} dx & dy & dx^2 & dy^2 & 1 \end{bmatrix}^T$ for a quadratic deformation vector computed from the relative offset of locations p^i and p^j. We can interpret $\beta^{ij}_{t^i, t^j}$ as a quadratic spring model that switches between a collection of springs tailored for a particular pair of mixtures (t^i, t^j). Because the spring depends on the mixture components, spatial constraints are dependent on local appearance. As in [5], we find this useful for modeling self-occlusions due to changes in viewpoint. The last element of $\beta^{ij}_{t^i, t^j}$ defines a "prior" or bias over which pair of mixtures should be selected.

Given a test image with an annotated bounding box, we find the maximum scoring part arrangement p and mixture assignment t that overlaps the bounding box by at least 50%. When E is tree-structured, this solution can be computed with dynamic programming [28]. The edge structure E is learned from the landmark annotations using maximum likelihood, as in [5]. Model parameters α, β are learned using a structural SVM, as in [28].

6. Experiments

Dataset: Our goal is to predict the functionally-important regions on the object in the form of segmentation masks. We use the publicly available dataset of [3], which annotates the 2009 PASCAL trainval set with keypoint annotations. We select 300 images for each of 5 object categories, intentionally ignoring images with severe occlusions. We randomly split this into equal-sized training and test sets. We define functional regions as such: `chairs` and `sofas` labeled with seat and backrest regions, `tv monitors` labeled with screens that people look at, `bicycles` labeled with the left and right handlebar and the seat, and `bottles` labeled with grasping regions and regions that one places their mouth on. For ease of annotation, we represent segmentation masks for regions as 4-sided polygons, as shown in Fig.2.

Evaluation: To avoid conflating issues of detection with functional region prediction, we assume we are given a test image with a bounding box around the object of interest as well as a object class label. This is similar to the protocol followed by the Action Classification and Person Layout challenges in the Pascal benchmark [8]. We use one of aforementioned models to predict functional segmentation masks (with affordance labels) on each test bounding

box. We evaluate our prediction by thresholding the area of intersection/union between a pair of (ground truth, predicted) masks with corresponding labels. For objects with multiple functional regions, we require ratios for **all** pairs of (ground truth, predicted) masks to be greater than a given threshold. We argue that, in order to correctly use a bicycle, one must *simultaneously* place both their hands and bum on the correct functional regions.We compare our aforementioned models: blind **LinReg** (1), **SpatialPrior** (2); bottom-up **SurfaceIndoor** (4), **Prior+Surface** (7); and top-down **NN** (8), **LatentDPM** [12], **FunctionalDPM** (9). Qualitative results are shown in Fig. 6, and quantitative results are shown in Fig. 7.

Region prediction: For `chair`, `sofa` and `bicycle`, both FunctionalDPM and Nearest Neighbor do better than all other baselines, particularly at higher overlap thresholds. Qualitative results (Fig. 6) suggest that high overlaps are needed, for say, a humanoid robot to parse an chair accurately enough to sit on it. Using a 50% overlap threshold, we find that top-down methods tend to accurately parse 5% of bicycles, 15% of chairs, and 40% of sofas. All methods, including blind baselines, tend to process bottles and tv-monitors equally well (with 25% and 90% correctly-parsed, respectively). Blind baselines do well because there is little shape variation in such objects, as suggested by their prior masks in Fig. 3. Bicycles are particularly challenging because handler bars and seats are fairly small, and so require precise localization to satisfy the overlap criterion. In some sense, this is indicative of true functional difficulty; its harder to ride a bike than sit on a sofa! Both NN and FunctionalDPM tend to consistently outperform the well-known LatentDPM baseline. The latter suggests that functional part-labeling is important, while the former suggests that structurally-rich models (with many mixtures and/or parts) maybe needed for accurate function prediction. The inferior performance of Surface Indoor [17], compared to our blind baselines, is surprising. We attribute this to a lack of encoding of object-level spatial structure. This is corroborated by the fact that it does significantly better when combined with a spatial prior (Prior+Surface).

Landmark prediction: We also evaluate the accuracy of various models in predicting functional landmarks. We define a landmark as correctly localized when its predicted location sufficiently overlaps with the ground-truth location, similar to the probability of correct part (PCP) measures used for pose estimation.We posit that functional region prediction and landmark prediction should correlate well, since they both are capturing object function. We show quantitative results for various models in Fig. 8. We measure a model's performance by plotting the percentage of test images for which a minimum number of keypoints on the object was correctly localized. The ordering of various models is consistent with what we observe in Fig. 7.

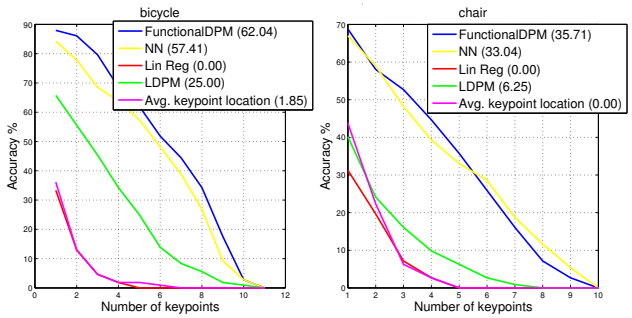

Figure 8. We plot the percentage of test images (Y-axis) that correctly localize a minimum number of landmarks on the object (X-axis), where the numbers in parentheses indicate percentage of test instances for which half the landmarks are correctly localized. We do not compare to bottom-up algorithms since they do not provide landmark predictions. We compare to an additional blind baseline that predicts the average landmark location, which sometimes outperforms blind regression. We see a similar trend as Fig. 7; top-down models such as NN and FunctionalDPM do quite well. Consistent results hold for other objects, but are omitted due to lack of space.

This further highlights one of the advantages of a landmark-based approach to modeling functional regions: we can leverage the large body of work in landmark prediction (say, of faces or articulated human poses). Or put another way, our functional perspective gives another motivation for predicting landmarks; instead of predicting expressions or articulated pose (typically limited to humans), one can predict general object *function*.

Conclusion: In this paper, we have revisited the idea of object affordances in the form of functionally-interesting regions on objects. We argue that functional regions should be placed alongside category labels, object segmentation masks, and attributes as desiderata that contemporary recognition systems should produce. We have shown how such regions can be represented in terms of affordance-labelled segmentation masks or functional landmarks. Finally, we have collected and annotated a general object dataset for initial exploration of this somewhat novel (yet classically-defined) problem. We evaluate a large collection of models, including simple "blind" baselines, existing bottom-up geometry-based techniques, and top-down shape-based models for this task. We show that top-down models that explictly reason about object shape and structure, encoded through functionally-supervised parts and non-parametric large-mixture models, are worthy of further exploration.

References

[1] H. Azizpour and I. Laptev. Object detection using strongly-supervised deformable part models. In *ECCV*, 2012. 2

[2] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010. 2

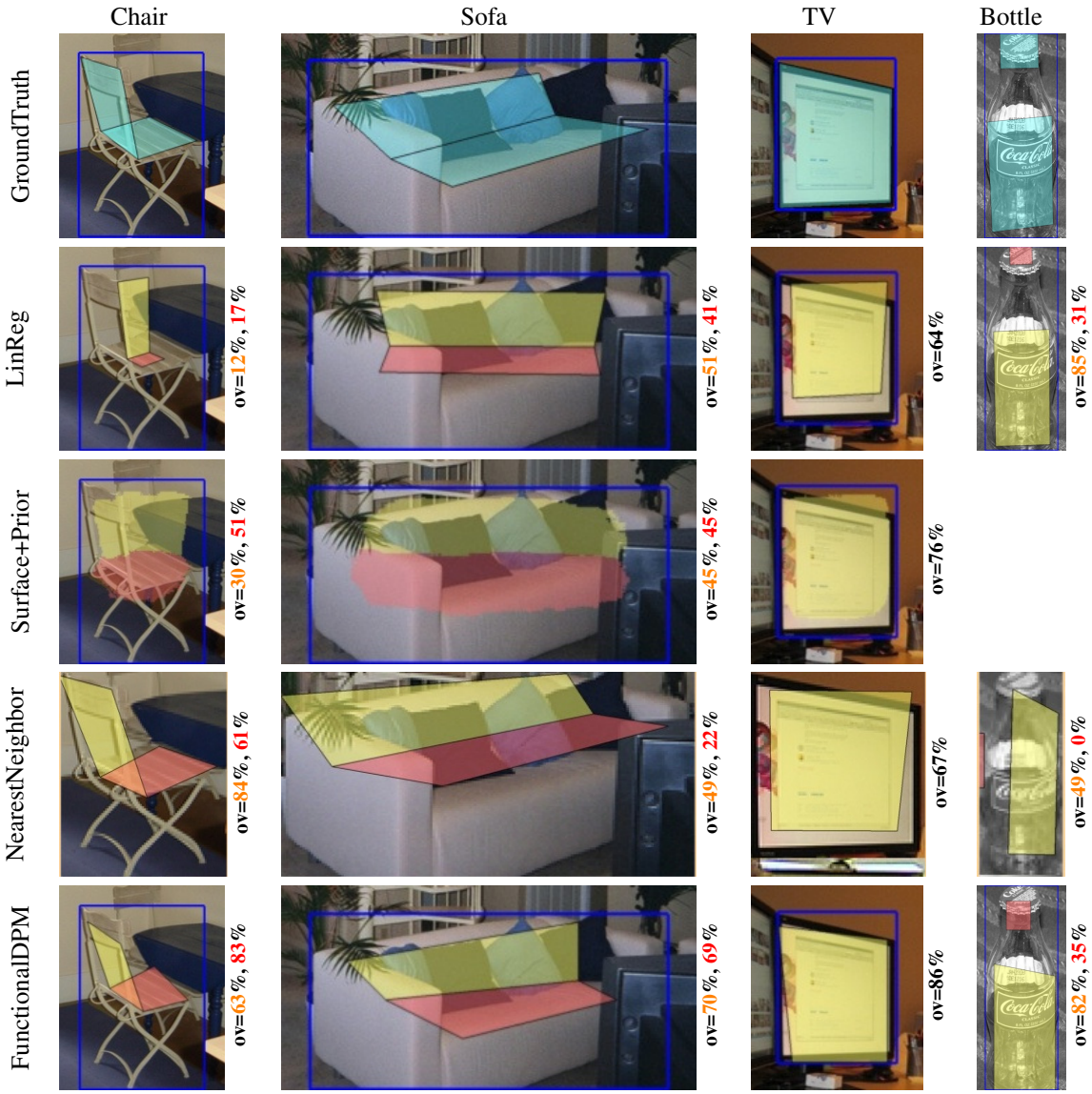

Figure 6. We compare model predictions versus ground-truth regions (shown as blue masks) for a test image from each object category. Yellow masks indicates the predicted backrest (for chairs and sofas), watchable monitor screens, and the graspable bottle regions. Red indicates predicted seats and bottle spouts. We compute overlaps percentages (using standard intersection over union [8]) for yellow and red predictions, displayed beside each image. "Good" predictions tend to overlap the ground truth by at least 50%. We qualitatively evaluate our models in such a manner in Fig. 7. Our bottom-up surface baselines do not generate geometric labels that are appropriate for curved objects such as bottles. In general, our top-down models perform better, but we refer the reader to the text for additional analysis.

[3] L. Bourdev, S. Maji, and J. Malik. Detection, attribute classification and action recognition of people using poselets (in submission). In *IEEE PAMI*. 2, 5

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*. IEEE, 2005. 5

[5] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*, 2012. 2, 5

[6] K. Duan, D. Parikh, D. J. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*. IEEE, 2012. 2

[7] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010. 1

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop, 2011. 5, 7

[9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*. IEEE, 2009. 1, 2

[10] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004. 3

[11] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005. 2

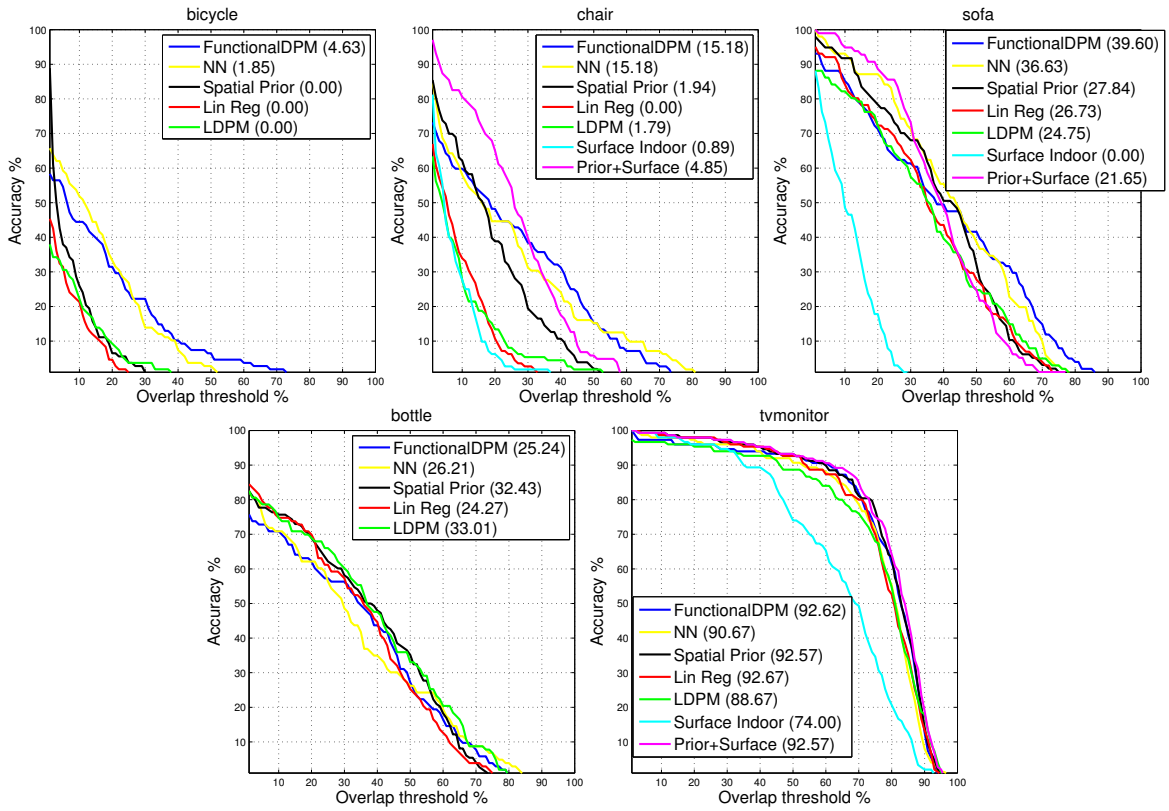

Figure 7. We plot accuracy of functional region prediction for a variety of models. We vary the functional region overlap threshold along the x-axis and compute the fraction of test images that satisfy the threshold (y axis). The numbers in parentheses indicate performance for 50% overlap. In general, top-down constraints are important for good performance. The geometric surface model of [17] does much better when combined with object-specific spatial models (Prior+Surface). NN and FunctionalDPM perform well for difficult categories such as the chair, sofa, and bicycle. Both models tend to outperform the latent DPM model of [12], indicating the importance large mixture-models and functional supervision. For categories with less within-class and viewpoint variation (such as bottle and tvmonitor), all models do well, including "blind" approaches that do not make use of pixel data (indicating the easiness of the problem in this case). Please see the text for further discussion.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 2010. 2, 4, 6, 8

[13] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic. People watching: Human actions as a cue for single-view geometry. In *ECCV*, 2012. 2

[14] J. Gibson. *The ecological approach to visual perception*. Hoiughton Mifflin, 1979. 1, 2

[15] H. Grabner, J. Gall, and L. Van Gool. What makes a chair a chair? In *CVPR*, pages 1529–1536. IEEE, 2011. 2

[16] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *(CVPR)*, 2011. 2

[17] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 2007. 2, 3, 4, 6, 8

[18] K. Koffka. Principles of gestalt psychology. 1935. 2

[19] H. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *International Journal of Robotics Research (IJRR)*, 2013. 2

[20] D. C. Lee, A. Gupta, M. Hebert, and T. Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. *NIPS*, 24, November 2010. 2

[21] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011. 2

[22] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *Computer Vision, 2009 IEEE 12th International Conference on*, 2009. 4

[23] E. Rivlin, S. Dickinson, and A. Rosenfeld. Recognition by functional parts. In *CVPR*, pages 267–274. IEEE, 1994. 2

[24] L. Stark and K. Bowyer. Achieving generalized object recognition through reasoning about association of function to structure. *IEEE PAMI*, 13(10):1097–1104, 1991. 2

[25] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *ICCV*, 2009. 2

[26] P. Winston, T. Binford, B. Katz, and M. Lowry. Learning physical descriptions from functional definition, examples, and precedents. In *In MIT press*, 1984. 2

[27] Y. Xiang and S. Savarese. Estimating the aspect layout of object categories. In *CVPR*. IEEE, 2012. 2

[28] Y. Yang and D. Ramanan. Articulated pose estimation using flexible mixture of parts. In *CVPR*. IEEE, 2011. 2, 5