Predicting the location of "interactees" in novel human-object interactions

Chao-Yeh Chen and Kristen Grauman

University of Texas at Austin

Abstract. Understanding images with people often entails understanding their *interactions* with other objects or people. As such, given a novel image, a vision system ought to infer which other objects/people play an important role in a given person's activity. However, while recent work learns about action-specific interactions (e.g., how the pose of a tennis player relates to the position of his racquet when serving the ball) for improved recognition, they are not equipped to reason about *novel* interactions that contain actions or objects not observed in the training data. We propose an approach to predict the localization parameters for "interactee" objects in novel images. Having learned the generic, action-independent connections between (1) a person's pose, gaze, and scene cues and (2) the interactee object's position and scale, our method estimates a probability distribution over likely places for an interactee in novel images. The result is a human interaction-informed saliency metric, which we show is valuable for both improved object detection and image retargeting applications.

1 Introduction

Understanding human activity is a central goal of computer vision with a long history of research. Whereas earlier work focused on precise body pose estimation and analyzed human gestures independent of their surroundings, recent research shows the value in modeling activity in the context of *interactions*. An interaction may involve the person and an object, the scene, or another person(s). For example, a person *reading* reads a book or paper; a person *discussing* chats with other people nearby; a person *eating* uses utensils to eat food from a plate. In any such case, the person and the **"interactee"** object (i.e., the book, other people, food and utensils, etc.) are closely intertwined; together they define the story portrayed in the image or video.

A surge of recent research in human action recognition aims to exploit this close connection [1–8]. Their goal is to improve recognition by leveraging human action (as described by body pose, appearance, etc.) in concert with the object being manipulated by the human. However, prior work is restricted to a closed-world set of objects and actions, and assumes that during training it is possible to learn patterns between a particular action and the particular object category it involves. For example, given training examples of *using a computer*, typical poses for typing can help detect the nearby computer, and vice versa; however,

Fig. 1. Despite the fact we have hidden the remainder of the scene, can you infer where is the object with which each person is interacting? Our goal is to predict the position and size of such "interactee" objects in a *category-independent* manner, without assuming prior knowledge of the specific action/object types.

in existing methods, this pattern would not generalize to help make predictions about, say, a person operating a cash register. Furthermore, existing work largely assumes that the interactions of interest involve a direct manipulation, meaning that physical contact occurs between the person and the interactee.

We seek to relax these assumptions in order to make predictions about novel, unseen human-object interactions. In particular, we consider the following question: *Given a person in a novel image, can we predict the location of that person's "interactee"—the object or person with which he interacts—even without knowing the particular action being performed or the category of the interactee itself?* Critically, by posing the question in this manner, our solution cannot simply exploit learned action-specific poses and objects. Instead, we aim to handle the open-world setting and learn generic patterns about human-object interactions. In addition, we widen the traditional definition of an interactee to include not only directly manipulated objects, but also untouched objects that are nonetheless central to the interaction (e.g., the poster on the wall the person is reading).

Why should our goal be possible? Are there properties of interactions that transcend the specific interactee's category? Figure 1 suggests that, at least for humans, it is plausible. In these examples, without observing the interactee object or knowing its type, one can still infer the interactee's approximate position and size. For example, in image 1.A, we may guess the person is interacting with a small object in the bottom left.

We can do so because we have a model of certain pose, gaze, and scene layout patterns that exist when people interact with a person/object in a similar relative position and size. We stress that this is without knowing the category of the object, and even without (necessarily) being able to name the particular action being performed. The ability to predict *where* an interactee object is independent of *what* it is would be valuable to vision systems that must analyze novel interactions from arbitrary categories.

Based on this intuition, our idea is to learn from data how the properties of a person relate to the interactee localization parameters. Given instances labeled with both the person and interactee outlines—from a variety of activities and objects—we train a probabilistic model that can map observed features of the person to a distribution over the interactee's position and scale. Then, at test

time, given a novel image and a detected person, we predict the most likely places the interactee will be found. Our method can be seen as an "interaction-informed" metric for object saliency: it highlights regions of the novel image most likely to contain objects that play an important role in summarizing the image's content.

The proposed approach addresses a number of challenges. They include designing a reliable data collection procedure to handle this somewhat unusual annotation task; developing a bank of descriptors to capture the "meta-cues" about human appearance that signal localized interactions; and presenting applications to exploit the interactee predictions. For the latter, we show that by focusing attention on regions in the image that are prominently involved in the human interaction, our method enables novel applications for priming object detectors and image retargeting. As we will see in Sec. 3.4, the ability to localize the object without categorizing it is precisely what enables these new tasks.

Our results on two challenging datasets, SUN and PASCAL Actions, demonstrate the practical impact. We show the advantages compared to an existing high-level "objectness" saliency method and a naive approach that simply looks for interactees nearby a person. Finally, we perform a human subject study to establish the limits of human perception for estimating unseen interactees.

2 Related Work

Human-object interactions for recognition A great deal of recent work in human activity recognition aims to jointly model the human and the objects with which he or she interacts [1–8]. The idea is to use the person's appearance (body pose, hand shape, etc.) and the surrounding objects as mutual context—knowing the action helps predict the object, while knowing the object helps predict the action or pose. For example, the Bayesian model in [2] integrates object and action recognition to resolve cases where appearance alone is insufficient, e.g., to distinguish a spray bottle from a water bottle based on the way the human uses it. Similarly, structured models are developed to recognize manipulation actions [9] or sports activities [4, 3] in the context of objects. Novel representations to capture subtle interactions, like playing vs. holding a musical instrument, have also been developed [5]. Object recognition itself can benefit from a rich model of how human activity [1] or pose [8] relates to the object categories. While most such methods require object outlines and/or pose annotations, some work lightens the labeling effort via weakly supervised learning [6, 7].

While we are also interested in human-object interactions, our work differs from all the above in three significant ways. First, whereas they aim to improve object or action recognition, our goal is to predict the location and size of an interactee—which, as we will show, has applications beyond recognition. Second, we widen the definition of an "interactee" to include not just manipulated objects, but also those that are untouched yet central to the interaction. Third, and most importantly, the prior work learns the spatial relationships between the human and object in an *action-specific* way, and is therefore inapplicable

to reasoning about interactions for any action/object unseen during training. In contrast, our approach is *action-* and *object-independent*; the cues it learns cross activity boundaries, such that we can predict where a likely interactee will appear even if we have not seen the particular activity (or object) before.

Carried object detection Methods to detect carried objects (e.g., [10, 11]) may be considered an interesting special case of our goal. Like us, the intent is to localize an interactee object that (in principle) could be from any category, though in reality the objects have limited scale and position variety since they must be physically carried by the person. However, unlike our problem setting, carried object detection typically assumes a static video camera, which permits good background subtraction and use of human silhouette shapes to find outliers. Furthermore, it is specialized for a single action (carrying), whereas we learn models that cross multiple action category boundaries.

Social interactions Methods for analyzing social interactions estimate who is interacting with whom [12–14], or categorize the type of physical interaction [15]. The "interactee" in our setting may be another person, but it can also belong to another object category. Furthermore, whereas the social interaction work can leverage rules from sociology [12] or perform geometric intersection of mutual gaze lines [13, 14], our task requires predicting a spatial relationship between a person and possibly inanimate object. Accordingly, beyond gaze, we exploit a broader set of cues in terms of body posture and scene layout, and we take a learning approach rather than rely only on spatial reasoning.

Object affordances Methods to predict object affordances consider an object [16, 17] or scene [18] as input, and predict which actions are possible as output. They are especially relevant for robot vision tasks, letting the system predict, for example, which surfaces are sittable or graspable. Our problem is nearly the inverse: given a human pose (and other descriptors) as input, our method predicts the localization parameters of the object defining the interaction as output. We focus on the implications for object detection and image retargeting tasks.

Saliency Saliency detection, studied for many years, also aims to make class-independent predictions about what is important in an image. While many methods look at low-level image properties (e.g., [19]), a recent trend is to *learn* metrics for "object-like" regions based on cues like convexity, closed boundaries, and color/motion contrast [20–23]. Such metrics are category-independent by design: rather than detect a certain object category, the goal is to detect instances of *any* object category, even those not seen in training. In contrast, methods to predict the relative "importance" of objects in a scene [24–26] explicitly use knowledge about the object categories present. Different from any of the above, our method predicts *regions likely to contain an object involved in an interaction*. We compare it extensively to a state-of-the-art objectness metric [21] in our experiments, showing the advantages of exploiting human interaction cues when deciding which regions are likely of interest.

3 Approach

To implement our idea, we learn probabilistic models for interactee localization parameters. In the following, we first precisely define what qualifies as an interactee and interaction (Sec. 3.1) and describe our data collection effort to obtain annotations for training and evaluation (Sec. 3.2). Then, we explain the learning and prediction procedures (Sec. 3.3). Finally, we briefly overview two example applications that exploit our method's interactee predictions (Sec. 3.4).

3.1 Definition of human-interactee interactions

First we must define precisely what a human-interactee[1] interaction is. This is important both to scope the problem and to ensure maximal consistency in the human-provided annotations we collect.

Our definition considers two main issues: (1) the interactions are not tied to any particular set of activity categories, and (2) an interaction may or may not involve physical contact. The former simply means that an image containing a human-object interaction of any sort qualifies as a true positive; it need not depict one of a predefined list of actions (in contrast to prior work [27, 28, 2–4, 6, 7]). By the latter, we intend to capture interactions that go beyond basic object manipulation activities, while also being precise about what kind of contact does qualify as an interaction. For example, if we were to define interactions strictly by cases where physical contact occurs between a person and object, then walking aimlessly in the street would be an interaction (interactee=road), while reading a whiteboard would not. Thus, for some object/person to be an interactee, the person ("interactor") must be paying attention to it/him and perform the interaction with a purpose.

Specifically, we say that an image displays a human-interactee interaction if either of the following holds:

1. The person is watching a specific object or person and paying specific attention to it. This includes cases where the gaze is purposeful and focused on some object/person within 5 meters. It excludes cases where the person is aimlessly looking around.
2. The person is physically touching another object/person with a specific purpose. This includes contact for an intended activity (such as holding a camera to take a picture), but excludes incidental contact with the scene objects (such as standing on the floor, or carrying a camera bag on the shoulder).

An image can contain multiple human-interactee relationships. We assume each person in an image has up to one interactee. At test time, our method predicts the likely interactee location for each individual detected person in turn.

3.2 Interactee dataset collection

Our method requires images of a variety of poses and interactee types for training. We found existing datasets that contain human-object interactions, like the

[1] An interactee refers to the thing a particular person in the image is interacting with; an interactee could be an object, a composition of objects, or another person.

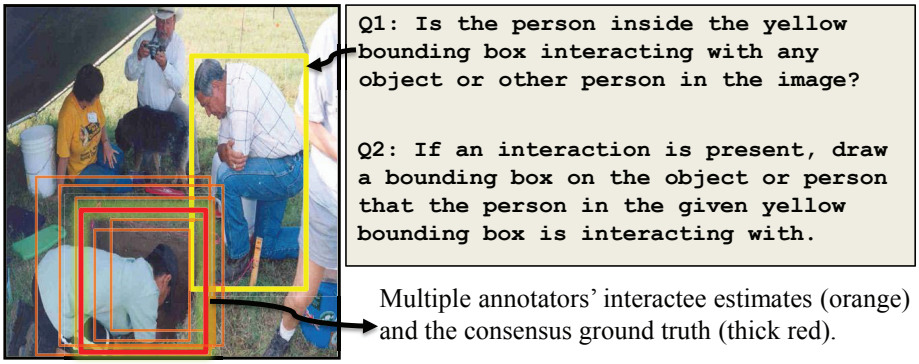


Fig. 2. Example annotation task. Top right shows (abbreviated) annotator instructions to identify the interactee for the person in the yellow bounding box. Here we also display in orange the boxes provided by 7 MTurkers, from which we compute the ground truth interactee (thick red box) as described in the text.

Stanford-40 and PASCAL Actions [27, 28], were somewhat limited to suit the category-independent goals of our approach. Namely, these datasets focus on a small number of specific action categories, and within each action class the human and interactee often have a regular spatial relationship. Some classes entail no interaction (e.g., *running*, *walking*, *jumping*) while others have a low variance in layout and pose (e.g., *riding horse* consists of people in fairly uniform poses with the horse always just below). While our approach would learn and benefit from such consistencies, doing so would essentially be overfitting, i.e., it would fall short of demonstrating action-independent interactee prediction.

Therefore, we curated our own dataset and gathered the necessary annotations. We use selected images from two existing datasets, SUN [29] and PASCAL 2012 [28]. SUN is a large-scale image dataset containing a wide variety of indoor and outdoor scenes. Using all available person annotations, we selected those images containing more than one person. The SUN images do not have action labels; we estimate these selected images contain 50-100 unique activities (e.g., *talking*, *holding*, *cutting*, *digging*, and *staring*). PASCAL is an action recognition image dataset. We took all images from those actions that exhibit the most variety in human pose and interactee localization—*using computer* and *reading*. We pool these images together; our method does not use any action labels. This yields a large number of unique activities.

We use Amazon Mechanical Turk (MTurk) to get bounding box annotations for the people and interactees in each image. The online interface instructs the annotators how to determine the interactee using the definition outlined above in Sec. 3.1. Figure 2 shows a condensed form; see Supp for more details. We get each image annotated by 7 unique workers, and keep only those images for which at least 4 workers said it contained an interaction. This left 355 and 754 images from SUN and PASCAL, respectively.

The precise location and scale of the various annotators' interactee bounding boxes will vary. Thus, we obtain a single ground truth interactee bounding box

via an automatic consensus procedure. First, we apply mean shift to the coordinates of all annotators' bounding boxes. Then, we take the largest cluster, and select the box within it that has the largest mean overlap with the rest.

The interactee annotation task is not as routine as others, such as tagging images by the objects they contain. Here the annotators must give careful thought to which objects may qualify as an interactee, referring to the guidelines we provide them. In some cases, there is inherent ambiguity, which may lead to some degree of subjectivity in an individual annotator's labeling. Furthermore, there is some variability in the precision of the bounding boxes that MTurkers draw (their notion of "tight" can vary). This is why we enlist 7 unique workers on each training example, then apply the consensus algorithm to decide ground truth. Overall, we observe quite good consistency among annotators. The average standard deviation for the center position of bounding boxes in the consensus cluster is 8 pixels. See Figure 5, columns 1 and 3, for examples.

3.3 Learning to predict an interactee's localization parameters

Training For each training image, we are given the bounding boxes for each person and its interactee. From each person box, we extract a descriptor $\boldsymbol{f} = [\boldsymbol{f}_p, \boldsymbol{f}_o, \boldsymbol{f}_s]$, composed of the following three features:

Body pose, $\boldsymbol{f}_p$: This feature captures the body pose of the person, which gives cues about how the person's posture and gesture relate to the interactee. For example, an extended arm may indicate that an interactee appears at the end of the reach; an extended leg may indicate an interactee is being kicked; a slouched torso may indicate holding a large object, while an upright torso may indicate holding a small light object, etc. We use a part-based pose representation called the *poselet activation vector* (PAV) [30]. A poselet is an SVM that fires on image patches with a given pose fragment (e.g., a bent leg, a tilted head). The PAV records how strongly each poselet is detected within the person bounding box. This yields a P-dim. vector for $\boldsymbol{f}_p$, where P is the number of poselets.

Orientation of head and torso, $\boldsymbol{f}_o$: These features capture the direction the person is looking or physically attending to. The head orientation is a proxy for eye gaze, and the torso orientation reveals how the person has situated his body with respect to an interactee. We predict both using the method of [30], which uses PAVs to train discriminative models for each of a set of discrete yaw intervals in $[-180°, 180°]$. This yields a 2-dimensional vector for $\boldsymbol{f}_o = [\theta_{head}, \theta_{torso}]$ consisting of the two predicted angles.

Scene layout, $\boldsymbol{f}_s$: This feature records the position of the person. It reflects the person in the context of the greater scene, which helps situate likely positions for the interactee. For example, assuming a photographer intentionally framed the photo to capture the interaction, then if the person is to the far right, the interactee may tend to be to the left. This yields a 2-dimensional vector for $\boldsymbol{f}_s = [X, Y]$, where X, Y denotes the normalized image position of the person.

The target "label" for $\boldsymbol{f}$ consists of the localization parameters for its interactee box: $\boldsymbol{\ell} = \langle (x, y), a \rangle$. The coordinates (x, y) specify the position, in terms of the vector from the person's center to the interactee's center. The area a specifies the size of the interactee. We normalize both components by the size of the per-

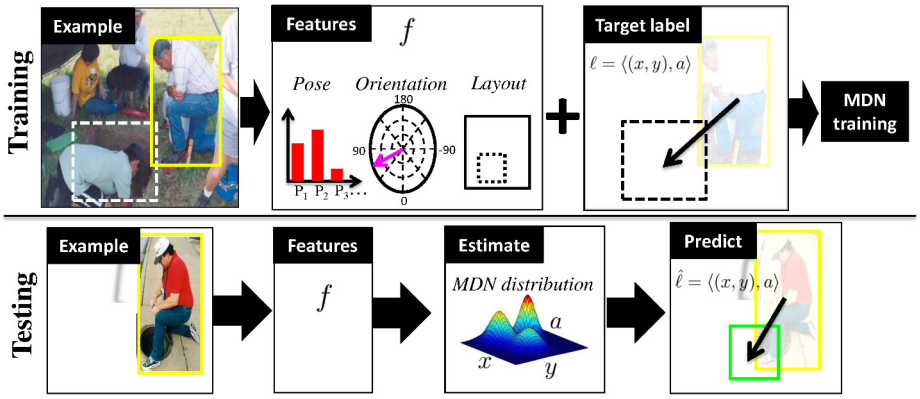


Fig. 3. Data flow in our approach. Top: training stage entails extracting features and target interactee positions/scales to learn a mixture density network (MDN). Bottom: testing stage entails estimating a mixture model from the learned MDN in order to predict the interactee's position/scale.

son (height plus width).[2] See Figure 3, top row. When there are multiple people in the training image, we record a set of features $\boldsymbol{f}$ for each one, separately, and pair it with that person's respective interactee label $\boldsymbol{\ell}$.

To build a predictive distribution for the interactee localization parameters, we want to represent a conditional probability density $P(\boldsymbol{\ell}|\boldsymbol{f}_k)$, for $k \in \{p, o, s\}$, where the subscript indexes the three features defined above. Since any given pose/gaze configuration may correspond to multiple feasible interactee localizations, we model this density as a mixture of Gaussians with m modes:

$$P(\boldsymbol{\ell}|\boldsymbol{f}_k) = \sum_{i=1}^{m} \alpha_i \mathcal{N}(\boldsymbol{f}_k; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{1}$$

where α_i denotes the prior mixing proportion for component i, $\boldsymbol{\mu}$ is its mean, and $\boldsymbol{\Sigma}_i$ is its covariance matrix.

Offline, we use the N labeled training examples $\{(\boldsymbol{f}^1, \boldsymbol{\ell}^1), \ldots, (\boldsymbol{f}^N, \boldsymbol{\ell}^N)\}$ to train a Mixture Density Network (MDN) [31] for each feature k. An MDN is a neural network that takes as input the observed features ($\boldsymbol{f}_k$), their associated parameters ($\boldsymbol{\ell}$), and the desired number of components m, and as output produces a network able to predict the appropriate Gaussian mixture model (GMM) parameters ($\boldsymbol{\alpha}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) for a novel set of observed features.[3]

We stress that our goal is to model interactions regardless of the type of the activity or the category of the interactee. Therefore, during training our method does not use any object or activity category labels.

Testing Given a novel test image represented by $\boldsymbol{f}^t$, our goal is to estimate the interactee's bounding box. First, we extract the descriptors from the person

[2] For this reason, it is not necessary to record scale in the scene layout feature above.

[3] We found it beneficial to model the two components of $\boldsymbol{\ell}$ with separate MDNs, i.e., one for position and one for area. Thus, altogether we have six MDNs, and predict $(\hat{x}, \hat{y})$ and $\hat{a}$ using their respective distributions in the test image.

bounding box in the novel image. Then, we use the learned MDN to generate the GMM $P(\boldsymbol{\ell}^t|\boldsymbol{f}_k^t)$ representing the most likely positions and scales for the target interactee. We get one GMM for each descriptor $\boldsymbol{f}_k$, where $k \in \{p, o, s\}$. Then, to fuse their predictions, we take the output of the model with the highest probability among all descriptors:

$$P(\boldsymbol{\ell}^t = \langle(\hat{x}, \hat{y}), \hat{a}\rangle \mid \boldsymbol{f}^t) = \max_{\boldsymbol{f}_k^t} P(\boldsymbol{\ell}^t \mid \boldsymbol{f}_k^t). \quad (2)$$

In this way, we can assign a probability to any candidate position and scale in the novel image.[4] To estimate the single most likely parameters $\boldsymbol{\ell}^*$ for $P(\boldsymbol{\ell}|\boldsymbol{f})$, we use the center of the mixture component with the highest prior (α_i), following [31]. The output interactee box is positioned by adding the predicted $(\hat{x}, \hat{y})$ vector to the person's center, and it has side lengths of $\sqrt{\hat{a}}$. See Figure 3, bottom row.

While all training images consist of true human-interactee interactions, it is possible a test image would have a human performing no interaction. In that case, the probabilistic outputs above can be used to reject as non-interactions those images whose interactee estimates are too unlikely.

3.4 Applications of interactee prediction

Our method is essentially an object saliency metric that exploits cues from observed human-interactions. Therefore, it has fairly general applicability. To make its impact concrete, aside from analyzing how accurate its predictions are against human-provided ground truth, we also study two specific applications that can benefit from such a metric.

Interactee-aware contextual priming for object detection First, we consider how interactee localization can prime an object detector. The idea is to use our method to predict the most likely place(s) for an interactee, then focus an off-the-shelf object detector to prioritize its search around that area. This has potential to improve both object detection accuracy and speed, since one can avoid sliding windows and ignore places that are unlikely to have objects involved in the interaction. It is a twist on the well-known GIST contextual priming [32], where the scene appearance helps focus attention on likely object positions; here, instead, the cues we read from the person in the scene help focus attention. Importantly, in this task, our method will look at the person (to extract $\boldsymbol{f}^t$), but will *not* be told which action is being performed; this distinguishes the task from the methods discussed in related work, which use mutual object-pose context to improve object detection for a particular action category.

To implement this idea, we run the Deformable Part Model (DPM) [33] object detector on the entire image, then we apply our method to discard the detections that are outside the 150% enlarged predicted interactee box (i.e., scoring them as $-\infty$). (To alternatively save run-time, one could apply DPM to only those windows near the interactee.)

[4] We also attempted a logistic regression fusion scheme that learns weights to associate per feature, but found the max slightly superior, likely because the confidence of each cue varies depending on the image content.

Interactee-aware image retargeting As a second application, we explore how interactee prediction may assist in image retargeting. The goal is to adjust the aspect ratio or size of an image without distorting its perceived content. This is a valuable application, for example, to allow dynamic resizing for web page images, or to translate a high-resolution image to a small form factor device like a cell phone. Typically retargeting methods try to avoid destroying key gradients in the image, or aim to preserve the people or other foreground objects. Our idea is to protect not only the people in the image from distortion, but also their predicted interactees. The rationale is that both the person and the focus of their interaction are important to preserve the story conveyed by the image.

To this end, we consider a simple adaption of the Seam Carving algorithm [34]. Using a dynamic programming approach, this method eliminates the optimal irregularly shaped "seams" from the image that have the least "energy". The energy is defined in terms of the strength of the gradient, with possible add-ons like the presence of people (see [34] for details). To also preserve interactees, we augment the objective to increase the energy of those pixels lying within our method's predicted interactee box. Specifically, we scale the gradient energy g within both person and interactee boxes by $(g + 5) * 5$.

4 Experimental Results

We evaluate four things: (1) how accurately do we predict interactees, compared to several baselines? (Sec. 4.1), (2) how well can humans perform this task? (Sec. 4.2), (3) does interactee localization boost object detection? (Sec. 4.3), and (4) does it help retargeting? (Sec. 4.4).

Baselines No existing methods predict interactee locations in a category independent manner. Therefore, to gauge our results we compare to the following three methods: (1) Objectness (Obj) [21], which is a state-of-the-art category-independent salient object detector. Like our method, it does not require information about the object category to detect; unlike our method, it does not exploit the interaction cues given by a person. We use the authors' code[5]. (2) Near person, which assumes that the interactee is close to the person. Specifically, it returns a bounding box centered at the person's center, with the same aspect ratio, and a size 40% of the person area (we optimized this parameter on training data). This is an important baseline to verify that interactee detection requires more sophistication than simply looking nearby the person. (3) Random, which randomly generates an interactee location and size.

Implementation details For each dataset, we use 75% of the data for training and 25% for testing, and resize images to 500×500 pixels. For both our method and Near person, we use the true person bounding boxes for both training and testing, to avoid conflating errors in interactee prediction with errors in person detection. When evaluating mAP, all methods consider sliding window candidates with a 25-pixel step size and 20 scales, and declare a hit whenever

[5] `http://groups.inf.ed.ac.uk/calvin/objectness/`

Metric	Dataset	Ours	Near person	Obj [21]	Random
Position error	SUN	**0.2331**	0.2456	0.4072	0.6113
	PASCAL	**0.1926**	0.2034	0.2982	0.5038
Size error	SUN	**33.19**	39.51	257.25	126.64
	PASCAL	34.39	**31.97**	206.59	100.31
mAP accuracy	SUN	**0.1542**	0.1099	0.0975	0.0450
	PASCAL	**0.1640**	0.1157	0.1077	0.0532

Table 1. Quantifying the accuracy of interactee localization with three metrics.

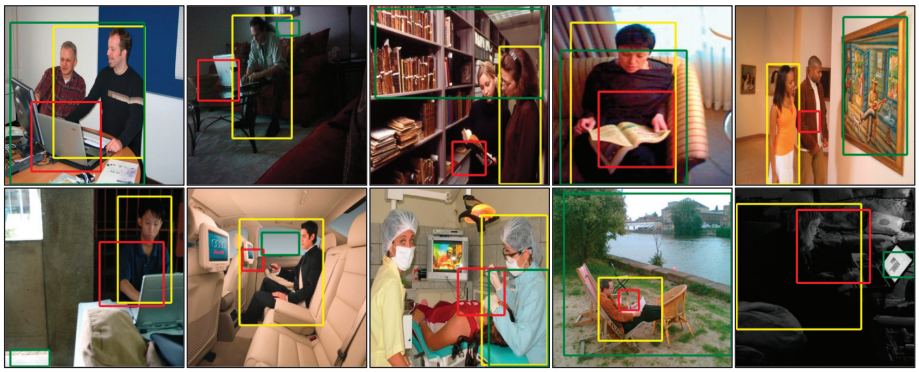


Fig. 4. Example interactee predictions for the given person (yellow), using our method (red) or Objectness [21] (green). Near person predicts a box centered at the person with $\sim$ 40% of its area (not shown for legibility). Note that there is no object detection involved in these predictions. Our method often accurately locates the interactee. Objectness can be distracted by the background or other objects (first four columns), while it works better than our method when the background is simple and the interactee is prominent (last column). Near person does not handle complex interactions well, but succeeds when the interactee is handheld and small (e.g., reading, 4^{th} column). Best viewed in color.

normalized overlap exceeds 0.3. Our method sorts the windows by their overlap with $\boldsymbol{l}^*$. For the MDNs, we use $m = 8$ and 10 mixture components on SUN and PASCAL, respectively, and use 10 hidden units. We use publicly available code[6] to compute the PAV vectors, which use $P = 1200$ poselets.

4.1 Accuracy of interactee localization

Table 1 compares the raw accuracy of interactee localization for all methods. We include three metrics to give a full picture of performance: position error, size error, and mean average precision (mAP). The errors are the absolute difference in position/area between the predicted and ground truth values, normalized by the person box size (height plus width) to prevent larger instances from dominating the result. The errors use only each method's most confident estimate (i.e., our $\boldsymbol{l}^*$, and the highest scoring box according to Obj and Near person).

[6] http://ttic.uchicago.edu/~smaji/projects/action/

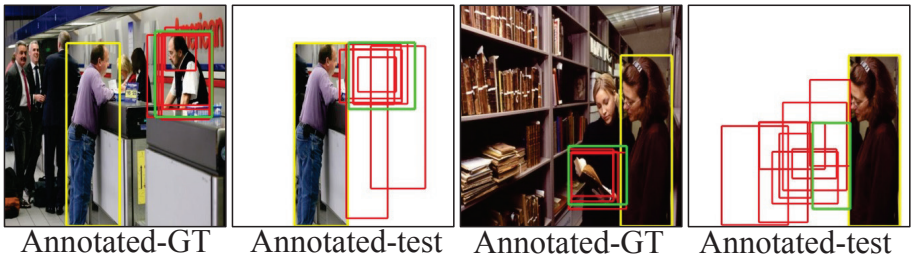


Fig. 5. We remove the background from the original image and ask human subjects to infer where the interactee might be. Red boxes denote their predictions, green box denotes consensus. Annotated-GT shows the full image (which is the format seen for ground truth collection, cf. Sec. 3.2). Annotated-test shows the human subject results. Naturally, annotators can more reliably localize the interactee when it is visible.

The mAP quantifies accuracy when the methods generate a ranked list of window candidates.

Our method outperforms the baselines on both datasets and all metrics, in all but one case. We improve average precision by 40% over the next competing baseline. Our error reductions on size and position are also noticeable.[7] The NEAR PERSON baseline does reasonably well, but suffers compared to our method because it is unable to predict interactees that don't entail physical contact, or those that rely on gaze and other high-level patterns (see last two rows of Table 1). It does, however, beat our method in terms of size error on PASCAL. Upon inspection, we find this is due to its easy success on the *reading* instances in PASCAL; people usually hold the reading material in their hands, so the interactee is exactly in the center of the body. We find OBJECTNESS suffers on this data by often predicting too large of a window covering much of the image. Our advantage confirms the value in making interaction-informed estimates of object-like regions. Figure 4 shows example predictions.

4.2 Human subject experiment

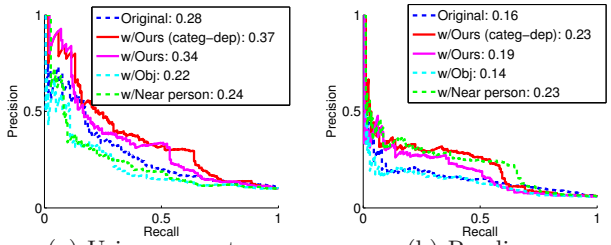
Next we establish an "upper bound" on accuracy by asking human subjects on MTurk to solve the same task. Our method localizes an interactee without observing the background content (outside of the person box) and without knowing what category the interactee belongs to. Thus, we construct an interface forcing humans to predict an interactee's location with a similar lack of information. Figure 5, columns 2 and 4, illustrate what the human subjects see, as well as the responses we received from 10 people.

Table 2 shows the human subjects' results alongside ours, for the subset of images in either dataset where the interactee is not visible within the person bounding box (since those cases are trivial for the humans and require no in-

[7] To help interpret the normalized errors: an error in predicted position of 0.20 amounts to being about 100 pixels off, while an error in predicted size of 33 amounts to about 6% of the image area.

	Human subject			Ours		
	Position error	Size error	mAP	Position error	Size error	mAP
SUN w/o visible	0.1573	28.92	0.3523	0.2736	36.58	0.1086
PASCAL w/o visible	0.0952	40.84	0.5226	0.2961	43.27	0.1750

Table 2. Results of the human subject test



(a) Using computer (b) Reading

Fig. 6. Interactee context helps focus the object detector. Numbers denote mAP.

ference).[8] The humans' guess is the consensus box found by aggregating all 10 responses with mean shift as before. The humans have a harder time on SUN than PASCAL, due to its higher diversity of interaction types. This study elucidates the difficulty of the task. It also establishes an (approximate) upper bound for what may be achievable for this new prediction problem.

4.3 Interactee-aware object detector contextual priming

Next we demonstrate the utility of our approach for contextual priming for an object detector, as discussed in Sec. 3.4. We use the PASCAL training images to train DPMs to find computers and reading materials, then apply our method and the baselines to do priming.

Figure 6 shows the results. We see our method outperforms the baselines, exploiting its inference about the person's attention to better localize the objects. While OURS uses action-independent training as usual, we also show a variant of our method where the MDN is trained only with images from the proper action class (see OURS (CATEG-DEP)). As expected, this further helps accuracy. Again, we see that NEAR PERSON fares well for the *reading* instances, since the book or paper is nearly always centered by the person's lap.

4.4 Interactee-aware image retargeting

Finally, we inject our interactee predictions into the Seam Carving retargeting algorithm, as discussed in Sec. 3.4. Figure 7 shows example results. For reference, we also show results where we adapt the energy function using OBJECTNESS's top object region prediction. Both methods are instructed to preserve the provided person bounding box. We retarget the source 500×500 images to 300×300.

[8] Since the test set here is a subset of the images, our numbers are not identical to our numbers in Table 1.

Fig. 7. Interactee-aware image retargeting example results. Our method successfully preserves the content of both the interactee (e.g., BBQ kit, book, painting of horse, laptop) and person, while reducing the content of the background. Objectness cannot distinguish salient objects that are and are not involved in the activity, and so may remove the informative interactees in favor of background objects. The bottom right example is a failure case for our method, where our emphasis on the interactee laptop looks less pleasing than the baseline's focus on the people. See Supp for more examples.

We see that our method preserves the content related to both the person and his interactee, while removing some unrelated background objects. In contrast, Objectness [21], unaware of which among the prominent-looking objects might qualify as an interactee, often discards the interactee and instead highlights content in the background less important to the image's main activity.

5 Conclusions

This work considers a new problem: how to predict where an interactee object will appear, given cues about a person's pose and gaze. While plenty of work studies action-specific object interactions, predicting interactees in an action-independent manner is both challenging and practical for various applications. The proposed method shows promising results to tackle this challenge. We demonstrate its advantages over multiple informative baselines, including a state-of-the-art object saliency metric, and illustrate the utility of knowing where interactees are for both contextual object detection and image retargeting. In future work, we are interested in exploring features based on more fine-grained pose and gaze estimates, and extending our ideas to video analysis.

Acknowledgements This research is supported in part by DARPA CSSG and a PECASE award from ONR.

References

[1] Peursum, P., West, G., Venkatesh, S.: Combining image regions and human activity for indirect object recognition in indoor wide-angle views. In: ICCV. (2005)

[2] Gupta, A., Kembhavi, A., Davis, L.: Observing human-object interactions: using spatial and functional compatibility for recognition. PAMI **31** (2009)

[3] Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for static human-object interactions. In: Workshop on Structured Models in Computer Vision, Computer Vision and Pattern Recognition (SMiCV). (2010)

[4] Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR. (2010)

[5] Yao, B., Fei-Fei, L.: Grouplet: A structured image representation for recognizing human and object interactions. In: CVPR. (2010)

[6] Ikizler-Cinbis, N., Sclaroff, S.: Object, scene and actions: Combining multiple features for human action recognition. In: ECCV. (2010)

[7] Prest, A., Schmid, C., Ferrari, V.: Weakly supervised learning of interactions between humans and objects. PAMI **34** (2012) 601–614

[8] Delaitre, V., Fouhey, D., Laptev, I., Sivic, J., Gupta, A., Efros, A.: Scene semantics from long-term observation of people. In: ECCV. (2012)

[9] Kjellstrom, H., Romero, J., Mercado, D.M., Kragic, D.: Simultaneous visual recognition of manipulation actions and manipulated objects. In: ECCV. (2008)

[10] Haritaoglu, I., Harwood, D., Davis, L.: W4: real-time surveillance of people and their activities. PAMI (2000)

[11] Damen, D., Hogg, D.: Detecting carried objects in short video sequences. In: ECCV. (2008)

[12] nd L. Bazzani, M.C., Paggetti, G., Fossati, A., Bue, A.D., Menegaz, G., Murino, V.: Social interaction discovery by statistical analysis of f-formations. In: BMVC. (2011)

[13] Marin-Jimenez, M., Zisserman, A., Ferrari, V.: Here's looking at you kid. detection people looking at each other in videos. In: BMVC. (2011)

[14] Fathi, A., Hodgins, J., Rehg, J.: Social interactions: a first-person perspective. In: CVPR. (2012)

[15] Yang, Y., Baker, S., Kannan, A., Ramanan, D.: Recognizing proxemics in personal photos. In: CVPR. (2012)

[16] Koppula, H., Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. In: RSS. (2013)

[17] Desai, C., Ramanan, D.: Predicting functional regions on objects. In: CVPR Workshop on Scene Analysis Beyond Semantics. (2013)

[18] Gupta, A., Satkin, S., Efros, A., Hebert, M.: From 3D scene geometry to human workspace. In: CVPR. (2011)

[19] Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. (2007)

[20] Liu, T., Sun, J., Zheng, N., Tang, X., Shum, H.: Learning to detect a salient object. In: CVPR. (2007)

[21] Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: CVPR. (2010)

[22] Endres, I., Hoiem, D.: Category independent object proposals. In: ECCV. (2010)
[23] Lee, Y.J., Kim, J., Grauman, K.: Key-Segments for Video Object Segmentation. In: ICCV. (2011)
[24] Spain, M., Perona, P.: Some objects are more equal than others: Measuring and predicting importance. In: ECCV. (2008)
[25] Hwang, S.J., Grauman, K.: Learning the relative importance of objects from tagged images for retrieval and cross-modal search. IJCV **100** (2012) 134–153
[26] Berg, A., Berg, T., Daume, H., Dodge, J., Goyal, A., Han, X., Mensch, A., Mitchell, M., Sood, A., Stratos, K., Yamaguchi, K.: Understanding and predicting importance in images. In: CVPR. (2012)
[27] Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L.J., Fei-Fei, L.: Action recognition by learning bases of action attributes and parts. In: ICCV. (2011)
[28] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV **88** (2010) 303–338
[29] Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN database: Large-scale scene recognition from abbey to zoo. In: CVPR. (2010)
[30] Maji, S., Bourdev, L., Malik, J.: Action recognition from a distributed representation of pose and appearance. In: CVPR. (2011)
[31] Bishop, C.M.: Mixture density networks. Technical report (1994)
[32] Torralba, A.: Contextual priming for object detection. IJCV **53** (2003) 169–191
[33] Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. PAMI **32** (2010) 1627–1645
[34] Avidan, S., Shamir, A.: Seam carving for content-aware image resizing. ACM Trans. Graph. **26** (2007) 10