Detecting Actions, Poses, and Objects with Relational Phraselets

Chaitanya Desai and Deva Ramanan

University of California at Irvine, Irvine CA, USA
{desaic,dramanan}@ics.uci.edu

Abstract. We present a novel approach to modeling human pose, together with interacting objects, based on compositional models of local visual interactions and their relations. Skeleton models, while flexible enough to capture large articulations, fail to accurately model self-occlusions and interactions. Poselets and Visual Phrases address this limitation, but do so at the expense of requiring a large set of templates. We combine all three approaches with a compositional model that is flexible enough to model detailed articulations but still captures occlusions and object interactions. Unlike much previous work on action classification, we do not assume test images are labeled with a person, and instead present results for "action detection" in an unlabeled image. Notably, for each detection, our model reports back a detailed description including an action label, articulated human pose, object poses, and occlusion flags. We demonstrate that modeling occlusion is crucial for recognizing human-object interactions. We present results on the PASCAL Action Classification challenge that shows our unified model advances the state-of-the-art for detection, action classification, and articulated pose estimation.

Action recognition is often cast as a k-way classification task; a person is either riding a bike, running, or talking on the phone, etc. For example, the PASCAL Action classification challenge requires one to label a human bounding-box (provided at test-time) with an action class. Such a formulation is limiting for two reasons. First, it assumes manual annotation of test data. In "real-world" unconstrained images, **detection** is crucial: how many people are riding a bike in this image, and where are they? Second, one may be interested in richer descriptions beyond a k-way class label. For instance, is this person riding a bike or about to mount it? Is he gripping the handlebar with one or both hands? Part of what makes this problem hard is that (1) humans can articulate and interact with objects in a variety of ways and (2) the resulting occlusions from those articulations and interactions are hard to model.

In this work, we present a novel approach to modeling human pose, together with interacting objects. Our model detects possibly multiple person-object instances in a single image and generates detailed spatial reports for each such instance. See Fig. 1 for an example of the output our model generates on a test image, *without* the benefit of any test annotation. Our approach unifies several recent lines of thought with classic models of human pose.

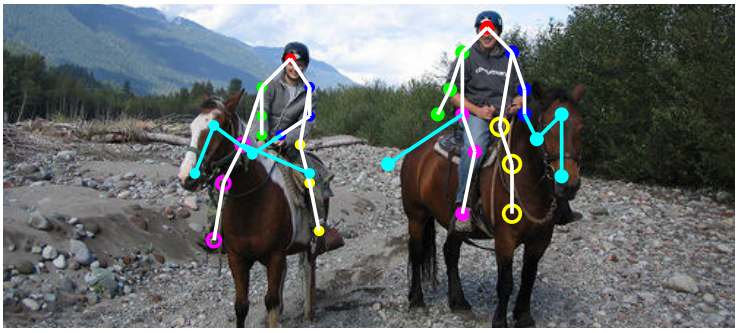


Fig. 1. Our model detects multiple people-objects, action class labels, human and object pose, and occlusion flag. The above result was obtained without any manual annotation of human bounding boxes at test-time. White edges connect human body parts. Light-blue edges connect object parts to each other and to the human. We define a *single* compositional model for each action class (in this case, `RidingHorse`) that is able to capture large changes in articulation, viewpoint and occlusions. We denote occluded parts by an open circle. For example, our model correctly predicts that a different leg of each rider is occluded behind his horse.

Articulated skeletons have dominated contemporary approaches for human pose estimation, popularized through 2D pictorial structure models that allow for efficient inference given tree-structured spatial relations [1]. We specifically follow the flexible mixtures of parts (FMP) framework of [2], which augments a standard pictorial structure with local part mixtures. While such methods are flexible enough to capture large variations in appearance due to pose, they still fail to accurately capture self-occlusions of limbs and occlusions due to interacting objects.

Visual phrases implicitly model occlusions and interactions through the use of a "composite" template that spans both a person and an interacting object [3]. Traditional approaches use separate templates for a person and object; in such cases, it may be difficult to model geometric and appearance constraints that arise from their interaction. Consider a person riding a horse; the person's legs tend to be occluded, while visible body parts tend to take on a riding pose. A single, global composite captures such constraints, but one may need a large number of composites to capture all possible person-horse interactions (a standing vs. galloping horse, an upright vs. crouched rider, etc.).

Poselets encode visual composites of parts rather than visual composites of objects [4]. A torso-arm composite implicitly captures interactions and occlusions that are difficult to model with separate templates for the arm and torso. By composing together different poselets, one can generate a large number of global composites. While such models are successful at detection, it is not clear if they can be used for detailed spatial reasoning, such as pose estimation. One reason is that a large number of poselets may be needed to capture all body poses. Another is that such methods lack a relational model that forces an anatomically-consistent arrangement of poselets to fire in a given detection.

Our approach combines the strengths of all three approaches. We break up global person+object composites into local patches or "phraselets," which can in turn be composed together to yield an exponentially-large set of composites. Notably, we enforce anatomically-consistent *relations* between phraselets to generate valid composites. We do so by defining phraselets as part mixtures in a FMP model, where local part mixture labels are obtained by "Poselet-like" clustering of global configurations of pose and nearby objects. To capture occlusions, we define separate phraselet mixtures for visible and occluded parts.

For example, we may learn different phraselets corresponding to hands gripping a handlebar, hands occluding torsos, and hands pointing away from the body. Our model includes relational constraints between phraselets; the presence of a handlebar phraselet induces a particular human body pose, as well as the presence of leg phraselets corresponding to legs occluded by bike-frames. Classic part models assume local appearance is independent of geometry; a hand looks the same regardless of the geometry of the remaining body. This makes occlusions and interactions difficult to model. Phraselets differ in that they encode dependencies between geometry and appearance through relational constraints.

Our model reports action class labels, articulated pose, object part locations, and part-occlusion flags. Notably, our models do *not* require a bounding-box annotation around a person at test-time. We show that our single model outperforms state-of-the-art methods for diverse tasks including visual composite detection (c.f. Visual Phrases), articulated pose estimation (c.f. FMP), and action classification (c.f. Poselets).

1 Related Work

Part models have a rich history in the context of pose-estimation. We refer the reader to a recent book chapter for a contemporary review [5]. Pictorial structures [1] are the dominant approach. Similar to [6], we learn part models in a discriminative framework. However, we follow a supervised learning framework for learning parts and relations, as in [2, 7]. Recent works have explored integrating relational part models with coarse-scale parts (rather than traditional limb models) [8]. This can also be integrated into a hierarchical, coarse-to-fine representation [9, 10]. Our model differs in that we consider only "fine" local representations, but focus on representing multi-modal appearances with mixture models. We show that one can represent an large set of coarse template by mixing and matching smaller patches. Our model jointly addresses detection, action classification, and pose estimation, similar to part-models that jointly reason about actions and pose [8], and detection and pose [9, 11].

Poselets were introduced and developed through [4, 12, 13]. We generate phraselets by clustering configurations of pose and nearby objects, unlike Poselets which clusters only pose. We consider action recognition as in [13], but also report human pose and object locations in a unified framework. Our phraselets differ in that they provide explicit reports of local occlusions. Perhaps most importantly, our model reports back an explicit articulated pose, while Poselets does not. Poselets are detected independently of each other, making it difficult to extract

Fig. 2. We show bike handles from PASCAL 2011 `RidingBike` action clustered using global configurations of pose and objects. Bike handles belonging to the same cluster are all assigned the same mixture label t^i as described in Sec. 2. Our clusters naturally encode changes in viewpoint, as well as different semantic object types; for example, the bottom-center and bottom-right clusters encode similar viewpoints, but different bicycle types (road bikes versus motorbikes). This is because each type induces different human poses, captured by our clustering algorithm.

a globally-consistent pose. Our relational model makes use of dynamic programming to force phraselets to fire in a globally-consistent manner.

Many approaches jointly recognize human pose and interacting objects. [14–16] describe contextual models for doing so, but assume that local part appearances are independent of the interaction. Such approaches typically assume a single instance of a person-object in the image. Our work differs in that we reason about multiple person-objects and detailed part occlusions of both the object and person. The latter allows us to better reason about occlusions arising from interactions. Visual phrases [3] takes a "brute-force" approach to modeling occlusions and pose interactions by defining a global template encompassing both the person and object. This approach may require a separate template for each combination of constituent objects and articulated pose. We instead use local mixtures and co-occurrence relations to reason about such interactions.

2 Phraselet clustering

We describe our approach for learning phraselets, or mixtures of local patches, specific to a given activity such as bike riding. We assume we are given images from an activity with keypoint labels spanning both the human body and any interacting objects. Typical keypoint labels may include *head, lt shoulder, rt elbow, lt ankle, etc* for the central figure and *front wheel, rear wheel, bike handle* for the bike. More details on the parts we collect keypoint locations for are given in Sec. 5. We assume these keypoint labels are with a visibility flag denoting if a particular keypoint is occluded or not.

(a) Visible elbow phraselets (b) Occluded elbow phraselets

Fig. 3. We show left-elbow phraselets learned from the `Running` action class in PASCAL VOC 2011. Our occluded clusters capture changes in the appearance of elbows resulting arising from viewpoint and occlusion.

Let $i \in \{1, 2, \ldots K\}$ be the one of the K parts of the person and/or the object specific to an activity. Let us write $p_n^i = (x, y)$ and $o_n^i \in \{0, 1\}$ for the pixel position and visibility flag of the i^{th} part in training image n, respectively. We write $t_n^i \in \{1, 2, \ldots M\}$ for a mixture or phraselet label. For the remainder of this section, we describe a method for obtaining mixture labels. Our intuition is that global changes in the geometric configuration of the human body and nearby object will produce local changes in appearance of a part i, and hence should be captured by t_i. For example, the local appearance of the hand will be affected by the orientation and type of bicycle (e.g., different bicycles can have different types of handlebars). We construct a feature vector associated with each part in each image, and cluster these vectors to derive mixture labels. To make the clustering scale invariant, we estimate a scale for each part in each image $s_n^i = \text{scale}_i * \text{headlength}_n$, where scale_i is the canonical scale of a part measured in human head-lengths, and head-length_n is the length of the head in image n. For example, we use $\text{scale}_i = 1$ for body parts and $\text{scale}_i = 2$ for bicycle wheels. We now write the feature vector for part i in image n as:

$$x_n^i = \left[\text{Dist Visible}\right]^T \tag{1}$$

$$\text{where} \quad \text{Dist} = \{w_{ij} d_{ij} : j = 1..K\}, \quad \text{Visible} = \{w_{ij} o_n^j : j = 1..K\} \tag{2}$$

$$\text{and} \quad w_{ij} = e^{-T_i ||d_{ij}||^2}, \quad d_{ij} = \frac{(p_n^j - p_n^i)}{s_n^i}$$

Dist is a 2K-vector of (weighted) pixel displacements of each of the K parts from part i, normalized for scale. $Visible$ is a K-vector of (weighted) binary occlusion flags. All terms are Gaussian-weighted by w_{ij} such that parts closer to part i have a larger influence in the global descriptor x_n^i. We found it useful to vary the variance of the gaussian (given by T_i) across each part, but use a fixed set across all activities For a given part i, we run K-means on all such features extracted from a training set of images.

Occlusion: Many parts are not visible in certain images. Such part instances may pollute a cluster if both visible and occluded parts are clustered together. Because we believe that occlusions will generate large changes in appearance, we simply separate x_n^i vectors into two sets, where part i is occluded or not, and separately run K means for each set. We generate $K = 6$ visible clusters and $K = 4$ occluded clusters for each part. This ensures that clusters/mixtures

1-6 are visible, while mixtures 7-10 are occluded. We show examples of visible clusters in Fig. 2. In Fig. 3, we compare visible and occluded clusters for the *left elbow* across images of people `Running`. We pad the image so that our model can find parts truncated by the image border; we treat truncation and occlusion identically, so that truncated patches along the border are added to the pool of occluded patches to be clustered.

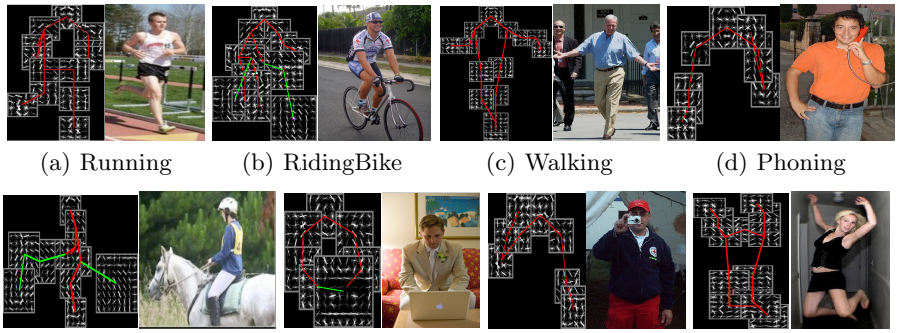
Relationship to past work: Our clustering algorithm is closely aligned to the Poselet clustering algorithm of [12], but with several key differences. Firstly, we consider the global configuration of the person *and* interacting object, rather than just the person. Secondly, we explicitly construct clusters corresponding to occluded parts. This allows us to generate such occlusion labels for detected parts at test time simply by reading off the estimated mixture label. Thirdly, and perhaps most importantly, our clusters consist of small patches that are forced to fire in globally-consistent arrangements, following a relational model described in the next section. This allows us to extract globally-consistent estimates of articulated poses. Our relational model also allows us to compose together a small number of phraselets with small spatial support into a large number of composites with large spatial support - we use roughly 100 template patches per activity, while Poselets requires roughly 1000 templates. One concern may be that phraselets are less discriminative than Poselets due to their small spatial support. However, a collection of phraselets can *learn* to behave like a single, larger Poselet by enforcing rigid relational constraints, as we show next.

3 Relational model

We now build an activity-specific model for scoring a collection of part mixtures, or phraselets. We would like to enforce consistent relations between phraselets, including spatial constraints on the geometric arrangement of parts, as well as appearance constraints on which mixtures can co-occur. Crucially, these constraints depend on each other; mixture appearance affects the spatial geometry and vice versa (e.g., a handlebar should be explained by an occluded phraselet only if the hand and handlebar lie spatially near each other). To encode such constraints, we follow the framework of [2], which describes a deformable part model that reasons about relations between local mixtures of parts. In this section, we review [2] and show how it can be used to build a relational model for phraselets.

Let I be an image, $p^i = (x, y)$ is the pixel location for part i and t^i is the mixture component of part i, derived from the previously described clustering algorithm. Suppose E is the edge structure defining relational constraints between the K parts. The score associated with a configuration of phraselets is written as

$$S(I, p, t) = b(t) + \sum_{i=1}^{K} \alpha^i_{t^i} \cdot \phi(I, p^i) + \sum_{i,j \in E} \beta^{ij}_{t^i, t^j} \cdot \psi(p^i - p^j) \tag{3}$$



(a) Running (b) RidingBike (c) Walking (d) Phoning

(e) RidingHorse (f) UsingComp. (g) TakingPhoto (h) Jumping

Fig. 4. Visualizations of our learned models and tree-structured relations. Our activity-specific tree connects part templates spanning both, the human and the object. Red edges connect parts of the human to each other. Green edges connect parts of an object to each other and to the human. Note that we are showing one (out of an exponential number of) combinations of local templates for each activity. For example, the selected phraselet mixtures in **(e)** correspond to a left-facing horse, but the same model generates other views by swapping out different mixtures at different spatial locations (as shown in Fig. 1).

Appearance relations: We write $b(t) = \sum_{ij \in E} w^{ij}_{t^i t^j}$ for a "prior" over mixture combinations, which factors into a sum of pairwise compatibility terms. This term might encode, for example, that curved handlebars tend to co-occur with road bicycles, while flat handlers tend to co-occur with motorbikes. Given that $\phi(I, p^i)$ is a feature vector (e.g., HOG [17]) extracted from pixel location p^i, the first sum from (3) computes the score of placing template $\alpha^i_{t^i}$, tuned for mixture t^i for part i, at location p^i.

Spatial relations: We write $\psi(p^i - p^j) = \left[dx\ dy\ dx^2\ dy^2\right]^T$ 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. For example, this dependency encodes the constraint that people may be posed differently for different types of bikes. Mixture-specific springs also encode self-occlusion constraints arising from viewpoint changes. For instance, our model can capture the fact that the right hip of a person is more likely to be occluded when it lies near a visible left hip, because such an spatial arrangement and mixture assignment is consistent with a right-facing person.

4 Inference and Learning

Inference corresponds to maximizing (3) with respect to p and t. When the relational graph is a tree, one can do this efficiently with dynamic programming, as described in [1, 2]. We omit the equations for a lack of space, but emphasize that our inference procedure returns back both part locations and part mixture

labels. While the inferred mixture labels in [2] are ignored, we use them to infer occlusion flags for each part.

Structure learning: Given a collection of K parts per activity, we would like to learn an activity-specific tree-based edge structure E connecting these K parts. The Chow-Liu algorithm is a well-known approach for learning tree models by maximizing mutual information [18, 1] of a given set of variables. In our case, we find the maximum-weight spanning tree in a fully connected graph whose edges are labeled with the mutual information between $z^i = (p^i, t^i)$ and $z^j = (p^j, t^j)$. Hence *both* spatial consistency and appearance consistency are used when learning the relational structure.

Once we learn an activity-specific tree, we learn the templates and relations for that tree using a structured prediction objective function. Let $z_n = \{(p_n^1, t_n^1)...(p_n^k, t_n^k)\}$ be a particular assignment of locations and types for all k parts in image n. Note that the scoring function in (3) is linear in the parameters $\theta = (\{w\}, \{\alpha\}, \{\beta\})$, and therefore can be expressed as $S(I_n, z_n) = \theta \cdot \Phi(I_n, z_n)$. We learn a model of the form:

$$\begin{aligned} & \underset{\theta, \xi_i \geq 0}{\operatorname{argmin}} && \frac{1}{2}\theta^T \cdot \theta + C \sum_n \xi_n \\ \text{s.t.} \quad & \forall n \in \text{positive images} && \theta \cdot \Phi(I_n, z_n) \geq 1 - \xi_n \\ & \forall n \in \text{negative images}, \forall z && \theta \cdot \Phi(I_n, z) \leq -1 + \xi_n \end{aligned} \tag{4}$$

The above constraint states that positive examples should score better than 1 (the margin), while negative examples, for all configurations of part positions and mixtures, should score less than -1. We collect negative examples from images consisting of people performing activities other than the one of interest. This form of learning problem is known as a structural SVM, and there exist many well-tuned solvers such as the cutting plane solver of SVMStruct in [19] and the stochastic gradient descent solver in [6]. We use the dual coordinate-descent QP solver of [2]. We show example models and their learned tree structure in Fig. 4 for 8 actions chosen from the PASCAL 2011 Action Classification competition.

5 Experiments

We consider 8 out of the 10 actions outlined in the PASCAL 2011 action classification competition. The actions considered correspond to the 8 models shown in Fig. 4. We train activity specific models as described in Sec. 4 for each of the 8 actions using PASCAL 2011-train data. In addition to the standard human joints, we model bike parts (*handle, front wheel, rear wheel*), horse parts (*nose, top-head, butt*) and computer screen (*whole object*) for their respective action classes. We model the full human body for most actions, but model only the upper body for actions with heavy occlusions and truncation (`Phoning`, `UsingComputer`, and `TakingPhoto`). We evaluate multiple aspects of our model, including detection, action classification, pose estimation, and occlusion prediction. To evaluate the latter two, we introduce novel evaluation schemes for evaluating poses under

Fig. 5. We show detection results obtained without *any* manual annotation of test images. We follow the notational conventions of Fig. 1, including open circles to denote occluded parts. Each row shows the N best detections for a single action model (denoted by the row's label). Our compositional models are able to capture large changes in viewpoint and articulation that are present even within a single action class.

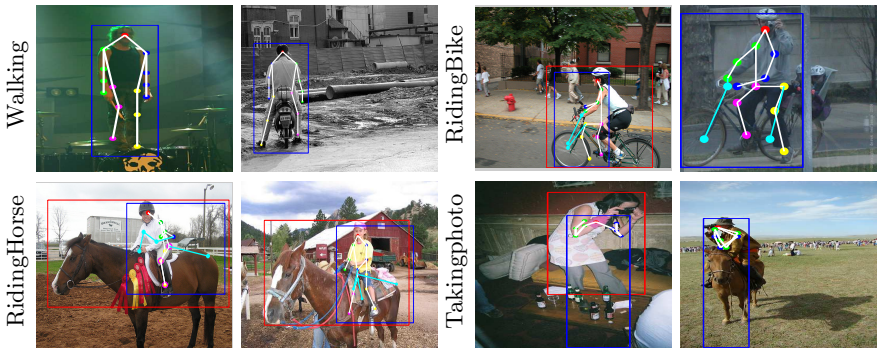


Fig. 6. We show 2 of the top false positives for a few actions. We plot ground-truth (red boxes) and predictions (blue boxes) belonging to only the action class denoted in each row. Many mistakes are due to imprecise bounding-box localization (`RidingHorse`) or confusion of action classes with similar poses (`Walking`). The latter is denoted by the lack of a red box. Some mistakes are due to inconsistencies and ambiguities in the ground-truth annotation. Consider the right image in the `RidingBike`/`TakingPhoto` rows; both images are annotated with a single action even though the person appears to be engaged in two actions (`TakingPhoto` and `RidingBike`/`RidingHorse`). This causes our predictions to be marked as false positives.

occlusion. Because the web-based PASCAL evaluation server is no longer evaluating entries on the 2011-test, we evaluate results on 2011-val. To do so, we have manually annotated both the train and val set with part locations and occlusion flags.

5.1 Action detection

For this task, our goal is to detect person-object composites in a test image. We use our models to produce composite candidates by running them as scanning-window detectors (without any manual annotation at test time), and applying NMS to generate a sparse set of non-overlapping detections. We visualize high scoring correct detections in Fig. 5 and false positives in Fig. 6. Ground truth person-object composites are obtained by considering a tight box around parts spanning the person and the object. To compare against groundtruth, we regress a rectangle using the part locations of the person and the object for each person-object detection.

We quantitatively evaluate our models using PASCAL's standard criteria of average precision (AP). We compare our models against a visual phrase (VP) baseline [3], trained for each action class. For those action classes without objects, this is equivalent to a standard DPM [6]. In both cases, we use defaults of 4 global mixtures and 6 parts per mixture. From Fig. 7, we see that our model outperforms these state-of-the-art baselines by a significant margin for most classes. The improvement is more modest for some classes (`Running`,`RidingBike`), perhaps because they exhibit less pose variation and so are well modeled by the global mixtures of the DPM.

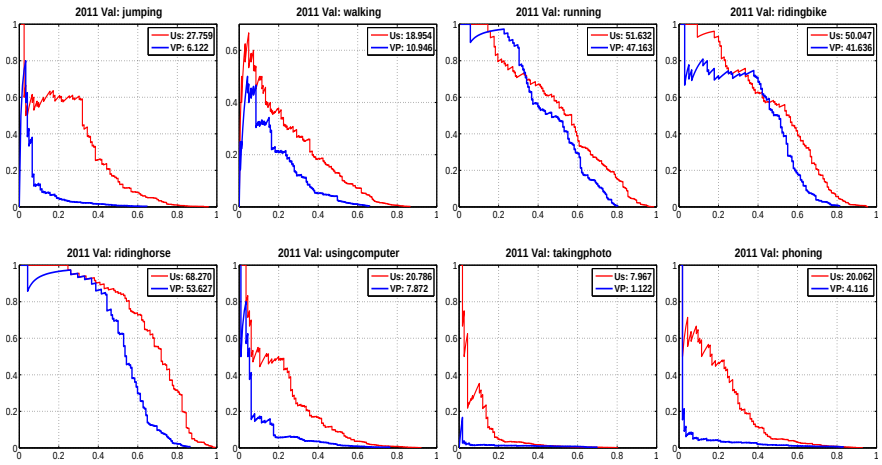


Fig. 7. Detection results on 2011 PASCAL-val set. Our model significantly outperforms a state-of-the-art visual phrase (VP) baseline [3].

5.2 Action classification

We compare our model against 2 other baselines apart from (VP/DPM): (1) FMP, the flexible articulated model of [2] applied to the joint person-object composite. (2) FMP+occ, which is obtained as follows: The FMP model estimates local mixtures by clustering the relative position of a part i wrt its parent j. FMP+occ also does this, but partitions the set of training data into visible/occluded instances of part i, and separately clusters each. This allows the FMP model to report visibility states using estimated part mixtures, analogous to our own model. To allow comparison to past work, we evaluate results following the protocol of PASCAL, assuming human bounding-boxes are given at test-time. We score each bounding box with the highest-scoring overlapping pose of each action model. For the (VP) baseline, we also give it access to a bounding box around the person-object composite. We present results on the 2011-val in Table 1. Our model outperforms state-of-the-art baselines, including DPM/VP on 7/8 actions. We also report numbers on 2010 test data using PASCAL's evaluation server, shown in Table 2 and compare to reported performance of [13]. Our numbers are comparable, even though [13] is trained using a large external dataset and includes additional post-processing steps (such as contextual re-scoring). Other state-of-the-art methods for action classification exist, but some may make intimate use of the annotated human bounding box on the test-image (say, to define a coordinate system to extract spatial features). We advocate action detection as a more realistic evaluation.

5.3 Person-object pose estimation

Qualitative results of our pose-estimation are shown in Fig 5. In general, our model rather accurately estimates parts of both the person and the object. Notably, our model also returns occlusion labels for each part (given by its estimated

Action classification on PASCAL 2011-val set

	Run.	R. Bike	R. horse	Phoning	TakingPhoto	UsingComp.	Walk.	Jump.
Us	69	81.7	90.3	32.9	24.3	45	40.3	49.6
FMP + occ	64.3	69.4	87.6	27.6	17.3	32.5	30.0	42.6
FMP	62.5	66.9	84.7	21.3	11.7	30.5	29.02	44.2
DPM/VP	63.2	66.4	79.7	21.2	12.1	43.5	32.1	28.8

Table 1. Class-specific AP results. In general our model strongly outperforms our baselines except for *UsingComp.* We suspect that this category exhibits less pose variation, and so is well-modelled by a global template.

Action classification on PASCAL 2010-test set

	Run.	R. Bike	R. horse	Phoning	TakingPhoto	UsingComp.	Walk.
Us	82.8	82.2	87.0	47.8	33.7	54.5	66.9
Poselets	85.6	83.7	89.4	49.6	31.0	59.1	67.9

Table 2. AP across various models on the PASCAL 2010 set. Our model is comparable to Poselets, even though the later is trained with a large external dataset and uses various post-processing steps for contextual res-coring.

mixture label). We quantitatively evaluate both aspects of pose estimation below.

Occlusion-aware pose evaluation: Standard benchmarks for pose estimation require an algorithm to report back the location of all parts, including those that may be occluded. See for example, the now-standard criteria of probability of a correct pose (PCP) [20]. We argue that a proper benchmark should only score visible parts. This is particularly relevant for human-object interactions because occlusions are rather common. We introduce a novel scheme for evaluating models and ground-truth poses that return a variable number of parts. Let n_g be the number of visible parts in the ground truth pose, and n_h be the number of visible parts in the hypothesized pose. Let k be the number of correctly matching parts across the two that are in correspondence and sufficiently overlap. We evaluate this pose using the fraction of correct parts $\frac{k}{.5(n_g+n_h)}$. One can show this is equivalent to the F_1 score, or harmonic mean of precision (the fraction of predicted parts that correctly match) and recall (the fraction of ground-truth parts that are correctly matched).

Results under our F_1 score are shown in Table 3. This evaluation penalizes algorithms for predicting an occluded part as visible; hence, it somewhat combines pose estimation with aspect estimation. Under this setting, our model outperforms all variants. The base FMP algorithm, like most algorithms for articulated pose estimation, reports a fixed set of parts. One may argue that it is artificially penalized under our F_1 score. However, FMP+occ *is* capable of predicting a visibility label per part, by construction, just as our model. We see that this model performs significantly better than FMP, but is still considerably lower than our final model.

We also score PCP in Table 4, which requires an algorithm to report locations of all parts, regardless of their visibility. Our algorithm still outperforms the 2 baselines. *This suggests our model accurately predicts the locations of even oc-*

Occlusion-aware F1 score

	Run.	R. Bike	R. horse	Phoning	TakingPhoto	UsingComp.	Walk.	Jump.
Us:	**66.8**	**49.2**	**65.3**	**41.4**	**30.8**	**41.2**	**44.8**	**40.3**
FMP+occ:	64.7	45	61.9	31.5	22.1	40.2	32.9	38.1
FMP	59.2	42.4	51.2	24.4	21.2	28.4	24.3	29.1

Table 3. Pose estimation across various models on 8 actions from PASCAL 2011, scored using F_1 score. The numbers reported are average F_1 scored over all test instances belonging to the action of interest. Algorithms are penalized for predicting the location of an occluded part in a test image. Our model outperforms state-of-the-art FMP model[2] by a significant margin, even when its augmented to encode occlusions.

PCP score

	Run.	R. Bike	R. horse	Phoning	TakingPhoto	UsingComp.	Walk.	Jump.
Us:	**68.7**	**50.7**	**64.7**	**39.9**	**28.9**	**43.1**	**45.4**	**40.7**
FMP+occ:	67.7	45.1	59.8	29.7	20.7	39.7	33.1	38.9
FMP	63.4	45.6	56.6	27.4	23.8	35.8	32.6	37.2

Table 4. Pose estimation across various models on 8 actions from PASCAL 2011, scored using PCP. Algorithms are required to predict the location of all parts (including occluded ones) in a test image. See text for details.

cluded parts. Interestingly, we still see a substantial improvement in performance from FMP to FMP+occ for most actions. In retrospect, this may seem obvious. Parts undergoing occlusions look different than when they are visible, and so one should train separate visual mixtures for such cases. One might suspect that these visual mixtures should have zero-weight, to ensure that no image evidence is scored during an occlusion. We take the view that the learning algorithm should determine this using training data. It may be that occluded parts still generate a characteristic gradient pattern (e.g., T-junctions), which can be captured by a template. Note that FMP+occ approach, in some sense, is a partial "phraselet" clustering since global knowledge of occluders is used to influence local appearance modeling.

Benchmark pose estimation: One could argue our phraselet model is directly applicable to pose estimation, without regard to interacting objects or actions. To evaluate this, we trained and evaluated our model on the PARSE benchmark [21]. We achieve a PCP score of 77.4%, outperforming the previous state-of-the-art FMP model at 74.9% (reported in [2]).

6 Conclusion

We have presented a novel approach to modeling human pose, together with interacting objects, based on compositional models of local visual interactions and their relations. Our modeling framework captures the complex geometry, appearance, and occlusions that arise in person-object interactions. We effectively use such models to detect person-object composites, estimate action class labels, articulated pose, object pose, and occlusion labels within a single, unified framework. We demonstrate compelling performance on diverse tasks includ-

ing detection, classification, and pose estimation, as evidenced by comparing to state-of-the-art models especially tuned for those tasks.

Acknowledgements: We thank Yi Yang for fruitful discussions, and Deepa Desai for graciously annotating images. Funding for this research was provided by NSF Grant 0954083, ONR-MURI Grant N00014-10-1-0933, and the Intel Science and Technology Center - Visual Computing.

References

1. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. IJCV (2005)
2. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR. (2011)
3. Sadeghi, M., Farhadi, A.: Recognition using visual phrases. In: CVPR. (2011)
4. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: ICCV. (2009)
5. Ramanan, D.: Part-based models for finding people and estimating their pose. In: Visual Analysis of Humans. (2011)
6. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE PAMI (2009)
7. Kumar, M., Zisserman, A., Torr, P.: Efficient discriminative learning of parts-based models. In: CVPR. (2010)
8. Yang, W., Wang, Y., Mori, G.: Recognizing human actions from still images with latent poses. In: CVPR. (2010)
9. Sun, M., Savarese, S.: Articulated part-based model for joint object detection and pose estimation. In: ICCV. (2011)
10. Wang, Y., Tran, D., Liao, Z. Forsyth, D.: Discriminative hierarchical part-based models for human parsing and action recognition. In: JMLR. (2012)
11. Tran, D., Forsyth, D.: Configuration estimates improve pedestrian finding. In: NIPS. (2007)
12. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting People Using Mutually Consistent Poselet Activations. In: ECCV. (2010)
13. Maji, S., Bourdev, L., Malik, J.: Action Recognition from a Distributed Representation of Pose and Appearance. In: CVPR. (2011)
14. Bangpeng, Y., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR. (2010)
15. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE PAMI (2009)
16. Prest, A., Schmid, C., , Ferrari, V.: Weakly supervised learning of interactions between humans and objects. IEEE PAMI (2011) in press.
17. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005)
18. Chow, C., Liu, C.: Approximating discrete probability distributions with dependence trees. Information Theory, IEEE Transactions on (1968)
19. Joachims, T., Finley, T., Yu, C.: Cutting plane training of structural SVMs. Machine Learning (2009)
20. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR. (2008)
21. Ramanan, D.: Learning to parse images of articulated bodies. In: NIPS. (2007)