Detecting Actions, Poses and Objects with Relational Phraselets

Chaitanya Desai
UC-Irvine, Irvine, CA, USA

Deva Ramanan
UC-Irvine, Irvine, CA, USA

Introduction: 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 [1] and Visual Phrases [4] 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. Fig 1 shows the tradeoffs associated with the three approaches. Our model brings together the strengths of all three approaches to solve detection, classification and detailed pose estimation using a **single** framework. See Fig. 2 for an example of the output our model generates on a test image. 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 (composite) detection, our model reports back a detailed description including an action label, articulated human pose, object poses, and occlusion flags. 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. An original version of this work appeared in [2]. Code and dataset will be available at http://www.ics.uci.edu/~desaic by the beginning of the conference.

Overview of our approach: We follow a supervised learning framework for learning parts and relations, as in [5, 3]. 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. Phraselets can be thought of as "parts" that straddle multiple objects. For example, a phraselet may model the hands of a person gripping the handlebar of a bicycle. Phraselets differ from traditional parts in one noteable aspect; their appearance varies drastically depending on the geometric arrangment of the objects; a handelbar looks quite different when a hand is not occluding it. To capture these constraints, we define phraselets as part mixtures inside an Flexible Mixture of Parts (FMP) framework like that of [5]. While [5] capture part appearances changes resulting from in-plane rotation by learning a mixture of part templates, our local "phraselet mixtures" are obtained by "Poselet-like" clustering of global configurations of pose of the human as well as the interacting object. 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 bikeframes. Thus, Phraselets encode dependancies between geometry and appearance through relational constraints.

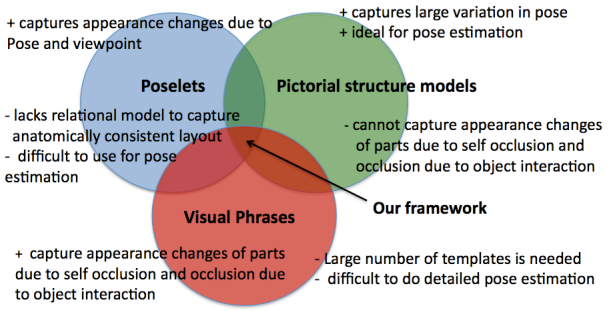

Figure 1. Our approach combines three distinct schools of thought inside a single unified framework, allowing us to simultaneously detect person-object composites , predict their action labels as well as produce detailed spatial pose for the person and the object

Fig 3 shows different phraselet clusters for image patches centered around bike handles, where our clustering is based on a "global pose-feature" that captures the overall pose of the person and the interacting object (in this case, the bike). This is very similar to the Poselet clustering approach of [1] . However, unlike Poselets, our clusters consist of small patches that are forced to fire in globally-consistent arrangments, following a relational model described in [5]. 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 compos-

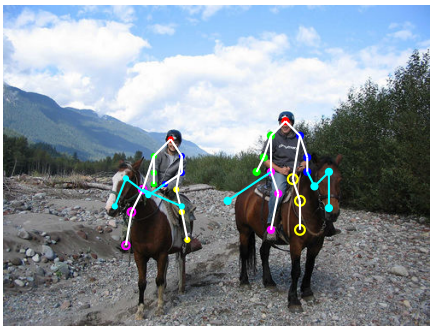

Figure 2. Our model detects multiple people-objects, action class labels, human and object pose, and occlusion flag. The above result was obtained without any manual annotation of human bounding boxes at test-time. White edges connect human body parts. Light-blue edges connect object parts to each other 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.

Figure 3. 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 inside the FMP framework of [5]. A single bike handle template is trained using patches belonging to the same cluster. Our clusters naturally encode changes in viewpoint, as well as different semantic object types;

ites with large spatial support - we use roughly 100 template patches per activity, while Poselets requires roughly 1000 templates.

Results: A visualization of 2 out of our learned 8 activity models is shown in 4. Results on Person object composite detection, Action Classification and pose estimation are shown in tables 1, 2 and 3respectively.

References

[1] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009. 1

[2] C. Desai and D. Ramanan. Detecting actions, objects and poses with relational phraselets. In *ECCV*, 2012. 1

[3] M. Kumar, A. Zisserman, and P. Torr. Efficient discriminative learning of parts-based models. In *CVPR*, 2010. 1

[4] M. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 1, 2

(a) RidingHorse (b) UsingComp.

Figure 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 **(a)** 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. 2).

Action Detection on PASCAL 2011-val

	Run.	R. Bike	R. horse	Phoning
Us	**51.3**	**50.0**	**68.2**	**20.0**
VP	47.1	41.6	53.6	4.1
	TakingPhoto	UsingComp.	Walk.	Jump.
Us	**7.9**	**20.7**	**18.9**	**27.7**
VP	1.1	7.8	10.9	6.1

Table 1. Detection results on 2011 PASCAL-val set. Our model significantly outperforms a state-of-the-art visual phrase (VP) baseline [4].

Action classification on PASCAL 2010-test set

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

PCP score

	Run.	R. Bike	R. horse	Phoning .
Us:	**68.7**	**50.7**	**64.7**	**39.9**
FMP	63.4	45.6	56.6	27.4
	Taking Photo	Using Comp.	Walk.	Jump
Us	**28.9**	**43.1**	**45.4**	**40.7**
FMP	23.8	35.8	32.6	37.2

Table 3. 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.

[5] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 1, 2