Human Pose as Context for Object Detection

Abhilash Srikantha[1,2]
abhilash.srikantha@tue.mpg.de

Juergen Gall[1]
gall@iai.uni-bonn.de

[1] University of Bonn,
Germany

[2] Max Planck Institute for Intelligent Systems,
Tuebingen, Germany

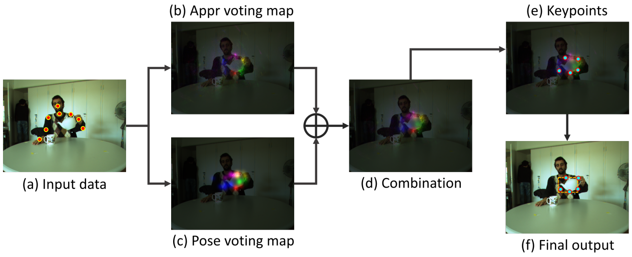

Figure 1: Detecting teapots: (a) Input is an image and automatically extracted human pose. (b) Object keypoint unaries based on appearance features and (c) using human pose features. (d) Linear combination of unaries. (e) Inferring keypoints (f) Regressing bounding box.

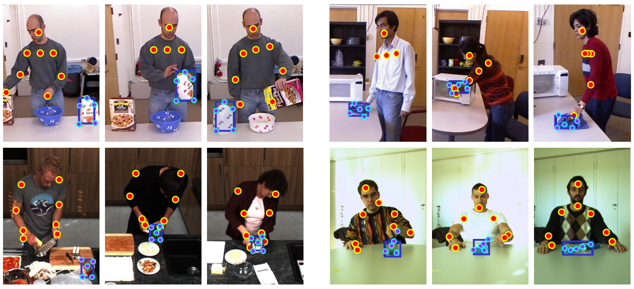

Figure 2: Qualitative results showing input human pose and most confident inferred bounding boxes as per Eqn (2). Successful detections are shown for classes *Milkbox, Cloth* from CAD-120; *Tin* from MPII-Cooking and *Roller* from ETHZ-Activity.

Object detection has seen considerable success, but the case of medium and small sized everyday objects still remains an open problem. Although such objects appear at low image resolutions, they often occur in the context of human interactions. Human context has been exploited in [1] which extends a deformable part model (DPM) to model spatial relations between body parts and parts of objects. This approach, however, only works well for images showing the instant of human-object interaction, i.e., when a human is closely in contact with an object. For images without an interaction, pose and objects are independently modelled, e.g., by having several models including either object or pose, or both together thereby leaving the human context unutilized.

In this work, we propose an approach that includes human pose as an additional context for object detection. Our approach is not limited to images showing explicit human-object interactions, but also works for general images where pose can be inferred. To this end, we model objects by a part based model and predict locations of parts from both image and pose data using regression forests. An outline of the approach is presented in Figure 1.

As illustrated in Figure 1(f), we represent an object by a set of descriptive keypoints $\mathcal{K} = \{\mathbf{k}_i\}$ where $\mathbf{k}_i$ encodes the image location of the i^{th} keypoint. Following pictorial structures model, an optimal keypoint configuration given an observation $\mathcal{D}$ is given by

$$p(\mathcal{K}|\mathcal{D}) \propto \prod_i \phi_i(\mathbf{k}_i) \cdot \prod_{i,j \in E} \psi_{ij}(\mathbf{k}_i, \mathbf{k}_j) \tag{1}$$

While we retain binary potentials to model relative keypoint offsets in the tree structured graph E as in [1], our work focuses on extracting more discriminative unary potentials $\phi_i(\mathbf{k}_i)$ derived from observations in appearance $\mathcal{D}_A$ and human pose $\mathcal{D}_P$ and is given by

$$\phi_i(\mathbf{k}_i) = p(\mathbf{k}_i|\mathcal{D}_A, \mathcal{D}_P) \tag{2}$$

$$= \left(K(\sigma_A) * \phi_i^A(\mathbf{k}_i)\right) + \alpha \left(K(\sigma_P) * \phi_i^P(\mathbf{k}_i)\right) \tag{3}$$

where $*$ represents the convolution operation and σ is the standard deviation for the Gaussian blur kernel K. Since the human pose can only provide a rough prior for the location of an object class but is insufficient for accurate object localization, $\sigma_P > \sigma_A$. Regressors based on appearance features are random forests base on image patches are similar to [3] with one main difference. Here, we do not scale normalize examples during training and the scale information from each patch is also stored at the leaves. Similarly, regressors based on human pose features are random forests that use extended features based on joint locations [6].

Table 1: average AUC measures for various Datasets.

Dataset	Appr.	Pose	Gall [3]	Desai [1]	Concat.	Comb.
MPII-Cooking	0.38	0.22	0.37	0.19	0.25	**0.41**
ETHZ-Action	0.46	0.24	0.42	0.50	0.23	**0.51**
CAD-120	0.31	0.09	0.29	0.21	0.20	**0.32**

$$\phi_i(\mathbf{k}_i, \hat{s}) = \sum_{m=1}^{M} \frac{1}{|\mathcal{T}_i|} \sum_{T \in \mathcal{T}_i} p_m(\mathbf{k}_i - \mathbf{j}_m | c, \hat{s}, L_T) \cdot p(c|L_T), \tag{4}$$

We evaluate the proposed approach on three datasets: ETHZ-Activity [2], CAD-120 [4] and MPII-Cooking [5]. Human pose is automatically inferred in all three datasets. We use the PASCAL-VOC measure for object detection. We compare our approach in various settings in Table 1. It can be seen that the appearance (Appr.) only features significantly outperform the pose (Pose) only features. In [3], Hough forests are used for object detection using a star model. When comparing it with our approach using appearance only features, we observe that the tree model is only slightly better than the star model.

The method [1] combines human pose estimation and object detection. The approach infact performs better than Pose only features in ETHZ-Action and CAD-120 datasets, but significantly worse in MPII-Cooking dataset. We therefore also implemented the approach for the MPII-Cooking dataset using random forests by using appearance based features and using the joints of the human pose as additional keypoints. The performance was comparable to [1] at 0.21 AUC.

As for combining appearance and pose features, we compare to an approach where is a single forest is trained on a concatenation of both features (Concat). The accuracy of this approach, however, drops sharply in contrast to appearance only features. Finally, combining both modalities as per Eqn (2) yields the best results in all three datasets with gains rangning from 1% to 5%.

[1] Chaitanya Desai and Deva Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*, pages 158–172. Springer, 2012.

[2] Juergen Gall, Andrea Fossati, and Luc Van Gool. Functional categorization of objects using real-time markerless motion capture. In *CVPR*, pages 1969–1976. IEEE, 2011.

[3] Juergen Gall, Angela Yao, Nima Razavi, Luc Van Gool, and Victor Lempitsky. Hough forests for object detection, tracking, and action recognition. *PAMI*, pages 2188–2202, 2011.

[4] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *IJRR*, pages 951–970, 2013.

[5] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, pages 1194–1201. IEEE, 2012.

[6] Angela Yao, Juergen Gall, and Luc Van Gool. Coupled action recognition and pose estimation from multiple views. *IJCV*, pages 16–37, 2012.