Discovering Object Functionality

Bangpeng Yao Jiayuan Ma Li Fei-Fei
Computer Science Department, Stanford University, Stanford, CA
{bangpeng,jiayuanm,feifeili}@cs.stanford.edu

Abstract

Object functionality refers to the quality of an object that allows humans to perform some specific actions. It has been shown in psychology that functionality (affordance) is at least as essential as appearance in object recognition by humans. In computer vision, most previous work on functionality either assumes exactly one functionality for each object, or requires detailed annotation of human poses and objects. In this paper, we propose a weakly supervised approach to discover all possible object functionalities. Each object functionality is represented by a specific type of human-object interaction. Our method takes any possible human-object interaction into consideration, and evaluates image similarity in 3D rather than 2D in order to cluster human-object interactions more coherently. Experimental results on a dataset of people interacting with musical instruments show the effectiveness of our approach.

1. Introduction

What is an object? Psychologists have proposed two popular philosophies of how humans perceive objects. One view asserts that humans perceive objects by their physical qualities, such as color, shape, size, rigidity, etc. Another idea was proposed by Gibson [15], who suggested that humans perceive objects by looking at their affordances. According to Gibson and his colleagues [14, 2], affordance refers to the quality of an object or an environment which allows humans to perform some specific actions. Recent studies [23] have shown that affordance is at least as important as appearance in recognizing objects by humans. An example is shown in Fig.1.

In the field of computer vision, while most previous work has emphasized modeling the visual appearances of objects [11, 3, 10], research on object/scene affordance (also called functionality[1]) is attracting more and more researchers' attention recently [17, 20, 16, 33, 13]. On the one hand, observing the functionality of an object (e.g. how humans interact with it) provides a strong cue for us to recognize the category of the object. On the other hand, inferring object functionality itself is an interesting and useful task. For example, one of the end goals in robotic vision is not to simply tell a robot "this is a violin", but to teach the robot how to make use of the functionality of the violin - how to play it. Further, learning object functionality also potentially facilitates other tasks in computer vision (e.g. scene understanding [4, 13]) or even the other fields (e.g. exploring the relationship between different objects [9]).

Figure 1. Humans can use affordance to perceive objects. In this image, although the violin is almost invisible, most humans can easily conclude this is an image of a human playing a violin based on the way the human interacting with the object.

In this work, our goal is to discover object functionality from weakly labeled images. Given an object, there might be many ways for a human to interact with it, as shown in Fig.2. As we will show in our experiments, these interactions provide us with some knowledge about the object

Figure 2. There are multiple possible modes of interactions between a human and a given object. Some interactions correspond to the typical functionality of the object while others do not.

[1]There are subtle differences between affordance and functionality in psychology. But in this paper, we use them interchangeably.

and hence reveal the functionalities of those objects. Furthermore, while inferring these types of interactions, our method also builds a model tailored to object detection and pose estimation for each specific interaction.

We propose an iterative model to achieve our goals. Using violin as an example, given a set of images of human-violin interactions, we discover different types of human-violin interactions by first estimating human poses and detecting objects, and then clustering the images based on their pairwise distances in terms of human-object interactions. The clustering result can then be used to update the model of human pose estimation and object detection, and hence human-violin interaction. Compared with previous human-object interaction and affordance work, we highlight the following properties of our approach:

- **Same object, multiple interactions:** Our method takes into account the fact that humans might interact with the same object in different ways, with only some typical interactions corresponding to object affordance, as shown in Fig.2. This differs from most previous approaches that assume a single type of human-object interaction for each object [17, 20].
- **Weak supervision:** Comparing with [18, 33], our method does not require annotations of human poses and objects on every training image. We only need a general human pose estimator and a weak object detector trained from a small subset of training images, which will be updated by our iterative model.
- **Unconstrained human poses:** Rather than being limited to a small set of pre-defined poses such as sitting and reaching [16, 13], our method does not have any constraint on human poses. This allows us to learn a larger variety of human-object interactions.
- **Bridging the gap between 2D and 3D:** Considering that the same human-object interaction might lead to very different 2D appearances (Fig.3) because of different camera angles from which the images are taken, we convert 2D human poses to 3D and then measure the similarity between different images. This allows us to obtain more semantically meaningful clusters as compared to previous work [33, 25].
- **Aiming for details:** The functionality we learn refers to the details of human-object interactions, e.g. the pose of the human, the object, as well as how the object should be used by humans. This makes our work different from most previous functionality work which mainly focuses on object detection [20, 25].

The rest of the paper is organized as follows. Sec.2 introduces related work, then Sec.3 elaborates on our approach of weakly supervised functionality discovery. Sec.4 demonstrates experimental results.

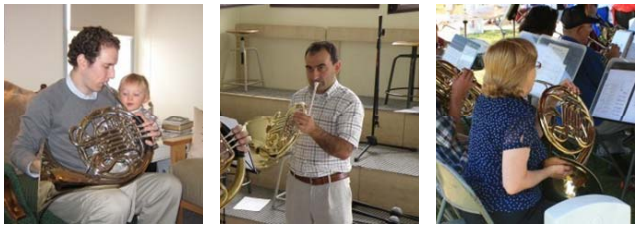

Figure 3. The same human pose might lead to very different appearances and 2D spatial configurations of body parts because of variations in camera angle.

2. Related Work

Functionality for object recognition. Recently, functionality has been used to detect objects [17, 20], where human gestures are recognized and treated as a cue to identify objects. In [16], 3D information is deployed such that one can recognize object affordance even when humans are not observed in test images. Such approaches assume that an object has the same functionality across all images, while our method attempts to infer object functionality given that humans might interact with the same object in many ways.

Human context. Context has been widely used in various computer vision tasks [26, 22]. Specifically, because of the advances in human detection [3, 10] and human pose estimation [1, 30], humans are frequently used as cues for other tasks, such as object detection (details below) and scene reconstruction [19, 4, 13]. Humans can also serve as context for each other to obtain performance improvement on all humans [6]. In this paper, we use human poses as context to discover functionalities of objects.

Human-object interaction. Our method relies on modeling the interactions between humans and objects. Most such approaches first estimate human poses [1, 30] and detect objects [10], and then model human-object spatial relationships to improve action recognition performance [5, 18, 33]. There are also approaches that directly train components of human-object interactions [8]. While those approaches usually require detailed annotations on training data, a weakly supervised approach is adopted in [25] to infer the spatial relationship between humans and objects. While our method also uses weak supervision to learn the spatial relationship between humans and objects, it takes into account that humans can interact with the same object in different ways, which correspond to different semantic meanings.

Semantic image clustering. In this paper, we use a clustering approach to discover different human-object interactions. Unsupervised learning of object classes has been explored in object recognition [28]. Recently, unsupervised object clustering [28] has been used to improve the performance of object classification. In this work, we cluster human action images in 3D, where the clustering results are more consistent with human perception than those from 2D.

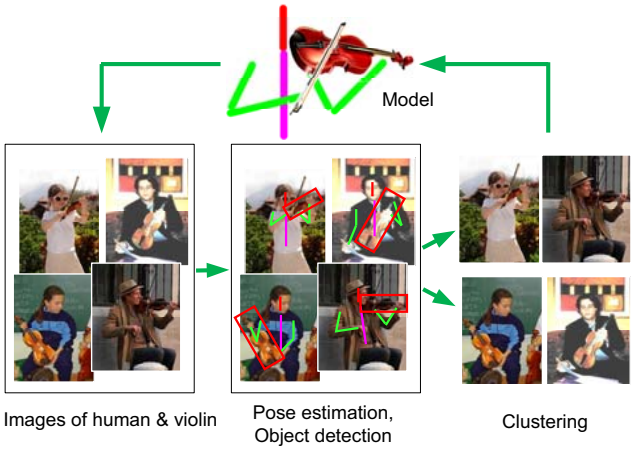

Figure 4. An overview of our approach ("violin" as an example). Given a set of images of human-violin interactions, our goal is to figure out what are the groups of interactions between a human and a violin, and output a model for this action.

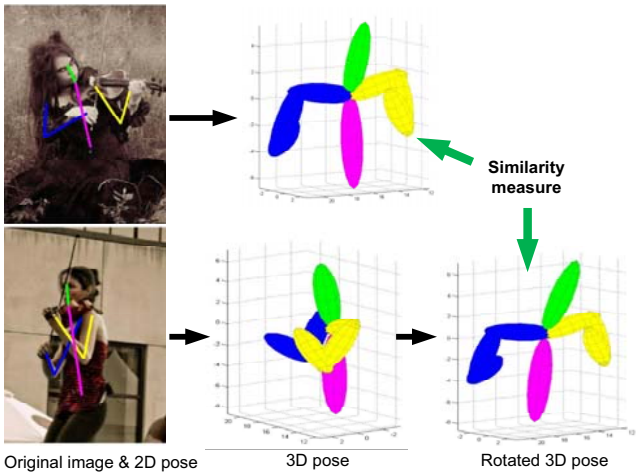

Figure 5. The pipeline we use to compute the similarity between two images of human-object interaction. We first detect objects and estimate human poses in each image, and then convert the key point coordinates in 3D and measure image similarity.

3. Algorithm

3.1. Overview

As shown in Fig.2, there are many possible ways for a human to interact with an object. Different interactions, such as playing a violin or using a violin as a weapon, correspond to different object functionalities. Our goal is to discover those interactions from weakly supervised data.

An overview of our approach is shown in Fig.4. Given a set of images of humans interacting with a certain object and an initial model of object detection and pose estimation, we propose an iterative approach to discover different types of human-object interactions and obtain a model tailored to each interaction. On the one hand, given a model of object functionality, we detect the object, estimate the human pose, convert 2D key points to 3D, and then measure the distance between each pair of images (Sec.3.2). The pairwise distance can then be used to decide which interaction type does each image belong to (Sec.3.3). On the other hand, given the clustering results, both the object detectors and human pose estimators can be updated so that the original model can be tailored to specific cases of object functionality (Sec.3.4).

3.2. Pairwise distance of human-object interactions

To reduce the semantic gap between human poses and 2D image representation (shown in Fig.3), we evaluate the pairwise distance of human-object interactions in the three-dimensional space. The pipeline we use to compute similarity between two images is shown in Fig.5. First, the 2D locations and orientations of objects and human body parts are obtained using off-the-shelf object detectors [10] and human pose estimation [30] approaches. Coordinates of 2D key points are then converted to 3D [27], and we evaluate pairwise distance between images by aligning 3D perspectives and computing the sum of squared distances between the corresponding body parts [32] and objects.

Object detection. We use the deformable parts model [10] to detect objects. To get more detailed information about human-object interactions, our detector also takes object orientation into consideration, as shown in Fig.4. At training time, we provide rectified bounding boxes with upright objects as positive training examples, and treat all the other image windows without the object or with non-upright objects as negative examples. During detection, we rotate the image using 12 different orientations and apply the trained detector in each case. Non-maximum suppression is done by combining the detection results on all orientations.

2D pose estimation. We use the flexible mixture-of-parts [30] approach for 2D human pose estimation. This approach takes the foreshortening effect into consideration, which facilitates the generation of 3D poses. We consider six body parts for the upper body of humans: head, torso, left/right upper arms, and left/right lower arms, as shown in Fig.5. For full-body humans, we also consider left/right upper legs and left/right lower legs. To improve performance, we replace the part filters with strong body-part detectors trained using the deformable parts model [10].

3D reconstruction of human pose. Because of camera angle changes, the same human pose might lead to very different 2D configurations of human body parts, as shown in Fig.3. Therefore, we use a data-driven approach to reconstruct 3D coordinates of human body parts from the result of 2D pose estimation [27]. By leveraging a corpus of 3D human body coordinates (e.g. CMU MOCAP), we recover

3D human body coordinates and camera matrices using a sparse linear combination of atomic poses. For the 3D locations of detected objects, we search the nearest body parts in 2D space, and average their 3D locations as the locations of objects in 3D space.

Pairwise distance computation. It has been shown that pose features perform substantially better than low-level features in measuring human pose similarities [12]. Following this idea and inspired by [32], we measure the distance of two human poses by rotating one 3D pose to match the other, and then consider the point-wise distance of the rotated human poses. Mathematically, let $\mathbf{M}_1$ and $\mathbf{M}_2$ be the matrices of the 3D key-point locations of two images $\mathbf{x}_1$ and $\mathbf{x}_2$. We find a rotation matrix $\mathbf{R}^*$ such that

$$\mathbf{R}^* = \arg\min_R \|\mathbf{M}_1 - \mathbf{M}_2\mathbf{R}\|^2, \tag{1}$$

and the similarity between $\mathbf{M}_1$ and $\mathbf{M}_2$ can be computed by $\mathcal{D}(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{M}_1 - \mathbf{M}_2\mathbf{R}^*\|^2$. We further incorporate the object in our similarity measure by adding the object as one more point in $\mathbf{M}$ and assuming that the depth of the object is the same as the hand that is closest to the object.

3.3. Clustering based on pairwise distance

The goal here is to cluster the given images so that each cluster corresponds to one human-object interaction, as shown in Fig.4. However, the task is not straightforward, since we only have the pairwise distance between images, rather than having a feature representation for each image.

We use an approach similar to spectral clustering [21] to address this issue. First, we use kernel PCA [29] to project each image $\mathbf{x}$ into a principal component space while keeping the pairwise image similarity computed from Sec.3.2. Denote the $N \times N$ similarity matrix as $\mathbf{K}$, where $\mathbf{K}_{ij} = \frac{1}{\mathcal{D}(\mathbf{x}_i,\mathbf{x}_j)+\epsilon}, \epsilon > 0$ is the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. Assuming an unknown feature representation for $\mathbf{x}_i$ as $\Phi(\mathbf{x}_i)$, we have the covariance matrix $\mathbf{C} = \frac{1}{N}\sum_{i=1}^{N}\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_i)^T$. Performing PCA, we have $\lambda_k\mathbf{v}_k = \mathbf{C}\mathbf{v}_k$, where λ_k is the k-th largest eigenvalue of $\mathbf{C}$. There also exist coefficients $\alpha_{k,1}, \cdots, \alpha_{k,N}$ such that

$$\mathbf{v}_k = \sum_{i=1}^{N}\alpha_{k,i}\Phi(\mathbf{x}_i). \tag{2}$$

Since $\mathbf{K}_{ij} = \Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j)$, the projection of $\Phi(\mathbf{x}_l)$ on $\mathbf{v}_k$ can be written as

$$z_{l,k} = \mathbf{v}_k^T\Phi(\mathbf{x}_l) = \sum_{i=1}^{N}\alpha_{k,i}\mathbf{K}_{il}. \tag{3}$$

According to [29], $\boldsymbol{\alpha}_k = [\alpha_{k,1}, \cdots, \alpha_{k,N}]^T$ can be computed by solving

$$N\lambda_k\boldsymbol{\alpha}_k = \mathbf{K}\boldsymbol{\alpha}_k, \quad s.t.\ \boldsymbol{\alpha}_k^T\boldsymbol{\alpha}_k = 1/\lambda_k. \tag{4}$$

Given the projected vector $\mathbf{z}_i$ for each image $\mathbf{x}_i$, we perform k-means clustering on all $i = 1, \cdots, N$ to form clusters of human-object interactions. Our approach chooses an appropriate number of clusters for every step of the process by using the standard elbow method - a cluster number is chosen such that adding another cluster does not give much decrement of the k-means objective. Since the above computation requires $\mathbf{K}$ to be positive semidefinite, we use a matrix approximation to replace $\mathbf{K}$ with $\hat{\mathbf{K}}$ such that

$$\hat{\mathbf{K}} = \arg\min \|\mathbf{K} - \hat{\mathbf{K}}\|^2, \tag{5}$$

where $\hat{\mathbf{K}}$ is positive semidefinite. We also assumed $\Phi(\mathbf{x}_i)$ to be centered in the above derivation. Please refer to [29] for details of how to drop this assumption.

3.4. Updating the object functionality model

In each iteration, we update the model of object detection and pose estimation for each cluster of human-object interaction. In each cluster, we re-train the models by using object detection and pose estimation results from this iteration as "ground-truth". Although there will be mistakes in these detection and estimation results, putting all the images together can still provide us more accurate priors that are tailored to each cluster.

In the step of object detection and pose estimation in the next iteration, we apply all the models from different clusters, and choose the one with the largest score of object detection and pose estimation. The detectors and estimators from different clusters are calibrated by fitting a probability distribution to a held-out set of images, as in [24].

4. Experiments

Dataset and experiment setup. For performance evaluation, we need a dataset that contains different interactions between humans and each object. The People Playing Musical Instrument (PPMI) dataset [31] contains images of people interacting with twelve different musical instruments: bassoon, cello, clarinet, erhu, flute, French horn, guitar, harp, recorder, saxophone, trumpet, and violin. For each instrument, there are images of people playing the instrument (PPMI+) as well as images of people holding the instrument with different pose, but not performing the playing action (PPMI-). We use the normalized training images to train our models, where there are 100 PPMI+ images and 100 PPMI- images for each musical instrument.

For each instrument, our goal is to cluster the images based on different types of human-object interactions, and obtain a model of object detection and pose estimation for each cluster. Ideally, images of humans playing the instruments should be grouped in the same cluster. To begin with, we randomly select 10 images from each instrument and annotate the key point locations of human body parts as

Instrument	Object detection		Pose estimation	
	Baseline	Ours	Baseline	Ours
Bassoon	16.4%	21.1%	43.1%	45.5%
Cello	41.9%	44.8%	48.1%	57.4%
Clarinet	11.1%	15.8%	52.0%	55.5%
Erhu	28.2%	33.1%	55.8%	57.8%
Flute	20.3%	23.1%	57.2%	59.7%
French horn	43.2%	43.7%	48.9%	55.1%
Guitar	45.5%	48.0%	40.8%	45.5%
Harp	30.6%	34.6%	41.0%	44.5%
Recorder	13.0%	16.9%	43.2%	51.5%
Saxophone	36.0%	41.9%	54.8%	60.7%
Trumpet	22.1%	24.7%	43.1%	48.6%
Violin	33.2%	39.5%	54.3%	63.5%
Overall	28.5%	32.3%	48.5%	53.8%

Table 1. Results of object detection and human pose estimation. "Baseline" indicates the results obtained by the original detectors [10] and the general pose estimator [30]. "Ours" indicates the results from the final models obtained from our iterative approach.

well as object bounding boxes, and train a detector [10] for each musical instrument and a general human pose estimator [30]. The object detectors and human pose estimator will be updated during our model learning process.

Object detection and pose estimation. Table 1 shows the results of object detection and pose estimation. For each musical instrument, we apply the "final" object detectors and pose estimators obtained from our method to the test PPMI images. For each image, we consider the models that correspond to the largest confidence score. We compare our method with the initial baseline models that are trained for all musical instruments. An object detection result is considered to be correct if the intersection of the result and the ground truth divided by their union is larger than 0.5, as in [7]. For human pose estimation, a body part is considered correctly localized if the end points of its segment lie within 50% of the ground-truth segment length [12].

The results show that our method outperforms the baselines by a large margin. This demonstrates the effectiveness of iteratively updating pose estimators and object detectors. Furthermore, our pose estimation result (53.8%) even performs slightly better than that in [33] (52.0%), where the models are trained with all PPMI training images annotated. The method in [33] (37.0%) obtains better object detection result than ours (32.3%), but was solving a simpler problem where object orientations were ignored.

Discovering object functionality. The PPMI dataset contains ground truths for which images contain people playing the instrument (PPMI+) and which images contain people only holding the instrument but not playing (PPMI-). This provides us the opportunity to evaluate the quality of clustering results. For each instrument, ideally, there exists a big cluster of humans playing the instrument, and many other clusters of humans holding the instruments but not playing. To get such clusters, we make use of the prior knowledge that there are 100 PPMI+ images for each instrument. We choose the number of clusters such that the number of images in the largest cluster is as close to 100 as possible. Fig.6 visualizes the average distribution of number of images in each cluster on all musical instruments.

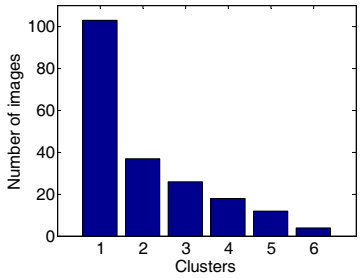

Figure 6. Average number of images per cluster on all musical instruments. The clusters are ordered by the number of images they contain.

We compare our clustering approach with two baselines. One is based on low-level image descriptors, where we represent an image with HOG [3] descriptors and then use PCA to reduce the feature dimension to 35, and then perform image clustering in the 35-dimensional space. In the other baseline, we cluster images based on 2D positions of keypoints of objects and human poses without converting them to 3D. For these two methods, we also choose the number of clusters on each instrument such that the number of images in the largest cluster is as close to 100 as possible.

For each instrument, we assume only the largest cluster contains images of people playing the instrument. A comparison of the accuracy of the different methods is shown in Fig.7. We observe that using 2D key points performs on par with low-level HOG features. The reason might be due to the errors in 2D pose estimation and the lack of accurate pose matching because of camera angle changes. On almost all the instruments, our method based on 3D key point locations significantly outperforms both low-level features and 2D key point locations. The only exception is on French horn, where all three approaches have similar performance. This is due to the large size of French horns, and the fact that the human poses as well as human-object spatial relationship are very similar in images of people playing French horn and people holding French horn but not playing. Finally, the performance can be further improved by combining 3D key points and low-level HOG features, as shown in Fig.7.

Affordance visualization. Examples of image clusters obtained by our approach are shown in Fig.8. On the instruments such as flute and trumpet, we are able to separate PPMI+ images from the others with high accuracy, because of the unique human poses and human-object spatial relationships on PPMI+ images. This partly explains why we

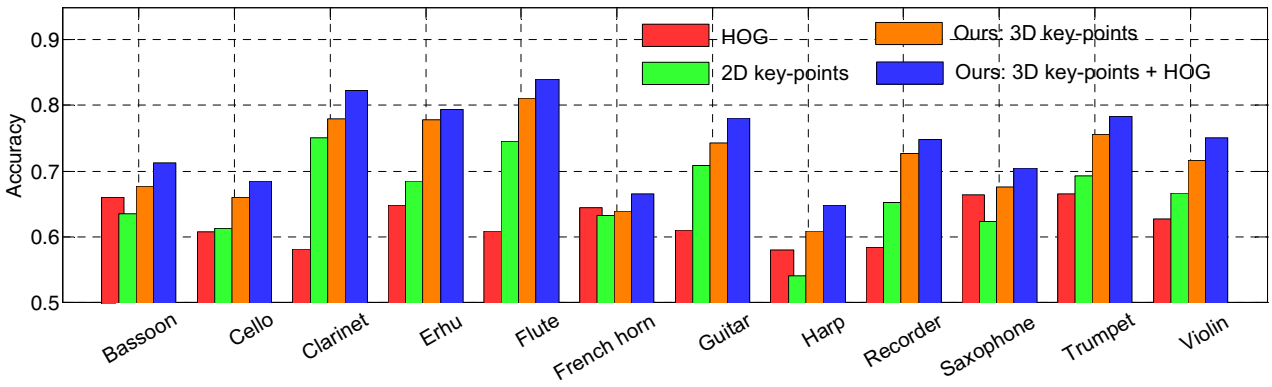

Figure 7. Comparison of our functionality discovery method with the approaches that based on low-level features or 2D key point locations.

Figure 8. Examples of image clusters obtained by our approach. For each instrument, images with the same border color belong to the same cluster. Solid lines indicate images of people playing the instrument (PPMI+) in the ground truth, while dashed lines indicate images of people holding the instrument but not playing it (PPMI-).

can obtain high accuracy on those instruments in Fig.7. The poor clustering performance on French horn can also be explained from this figure, where the spatial relationship between humans and French horns are very similar in images of all types of interactions.

Fig.9 visualizes the heatmap of the locations of human hands with respect to the musical instruments, as well as the locations of objects with respect to the average human pose in different interactions. On most instruments, we observe more consistent human hand locations on the clusters of people playing the instrument than that on the other clusters. However, we still observe some points that are frequently touched by the humans even for the cases of "holding but not playing" for some instruments, e.g. flute and guitar as shown in Fig.10. This shows some general rules when humans interact with a specific type of object, no matter what the functionality of the interaction is. Interestingly, people usually touch different parts of French horn when they are playing or not playing it, as shown in Fig.8.

Predicting objects based on human pose. Our method learns the interaction between humans and objects. Given a human pose, we would like to know what object the human is manipulating. On the PPMI test images, we apply all the human pose models to each image, and select the human that corresponds to the largest score. We say that the object involved in the selected model is manipulated by this

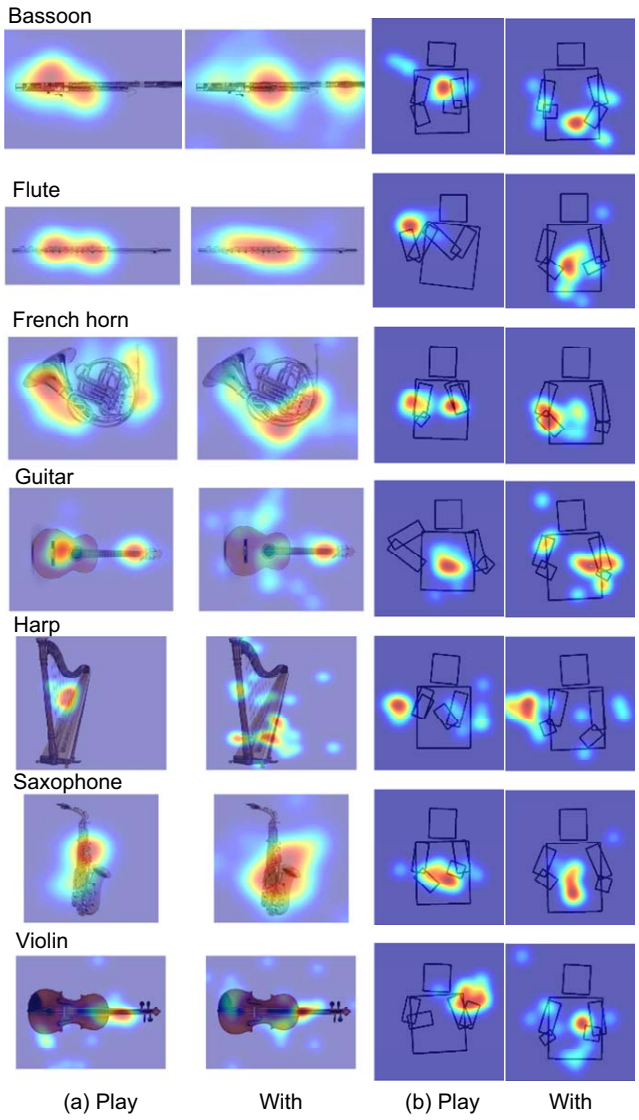

Figure 9. **(a)** Heatmaps of the locations of human hands with respect to musical instruments. **(b)** Heatmaps of the locations of objects with respect to the average human pose. For each instrument, "play" corresponds to the largest cluster, while "with" corresponds to all other clusters. We only show results from seven instruments due to space limitation. This figure is best viewed in color.

human. We only consider PPMI+ test images in this experiment. We compare our approach with a baseline that runs deformable parts models [10] of all instruments on each image, and output the instrument that corresponds to the largest calibrated score. The results are shown in Table 2.

Table 2 shows that on the musical instruments where the human pose is different from the others, such as flute and violin, our method has good prediction performance. On musical instruments which are played with a similar human pose, such as bassoon, clarinet and saxophone (shown in Fig.11), the appearance-based models perform better. This

Figure 10. Humans tend to touch similar locations of some musical instruments, even when they are not playing it.

Instrument	DPM	Ours	Instrument	DPM	Ours
Bassoon	**47%**	38%	Cello	39%	**49%**
Clarinet	32%	**38%**	Erhu	**53%**	23%
Flute	41%	**60%**	French horn	**78%**	37%
Guitar	**46%**	26%	Harp	51%	**53%**
Recorder	32%	**42%**	Saxophone	**53%**	29%
Trumpet	**59%**	53%	Violin	34%	**48%**

Table 2. Comparison of using appearance and using human pose to predict object categories. For each instrument, bold fonts indicate better results. Chance performance is 8%.

confirms that both object appearance and functionality are important in perceiving objects and provide complementary information [23].

5. Conclusion

In this paper, we propose a weakly supervised approach to learn object functionality, e.g. how humans interact with objects. We consider multiple possible interactions between humans and a certain object, and use an approach that iteratively clusters images based on object functionality and updates models of object detection and pose estimation. On a dataset of people interacting with musical instruments, we show that our model is able to effectively infer object functionalities. One direction of future research is to extend our

Figure 11. Humans might manipulate different objects with very similar poses.

method to the objects where the human-object interaction is more complicated, such as basketball.

Acknowledgement. We would like to thank Chris Baldassano, Abraham Botros, Jia Deng, and Vignesh Ramanathan for valuable comments to the paper. Li Fei-Fei is partially supported by an ONR MURI, an Intel ISTC-PC, and DARPA CSSG. Bangpeng Yao is partially supported by the Microsoft Research PhD Fellowship and SAP Stanford Graduate Fellowship.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009. 2

[2] L. Carlson-Radvansky, E. Covey, and K. Lattanzi. What effects on Where: Functional influence on spatial relations. *Psychol. Sci.*, 10(6):519–521, 1999. 1

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1, 2, 5

[4] V. Delaitre, D. F. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. A. Efros. Scene semantics from long-term observation of people. In *ECCV*, 2012. 1, 2

[5] V. Delaitre, J. Sivic, and I. Laptev. Learning person-object interactions for action recognition in still images. In *NIPS*, 2011. 2

[6] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. In *ECCV*, 2010. 2

[7] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. 5

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

[9] C. Fellbaum. WordNet: An electronic lexical database, 1998. 1

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminantly trained part-based models. *IEEE T. Pattern Anal. Mach. Intell.*, 32:1627–1645, 2010. 1, 2, 3, 5, 7

[11] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *ICCV*, 2003. 1

[12] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008. 4, 5

[13] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic. People watching: Human actions as a cue for single view geometry. In *ECCV*, 2012. 1, 2

[14] E. Gibson. The concept of affordance in development: The renascence of functionalism. *The Concept of development: The innesota Symp. on Child Psychology*, 15:55–81, 1982. 1

[15] J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979. 1

[16] H. Grabner, J. Gall, and L. V. Gool. What makes a chair a chair? In *CVPR*, 2011. 1, 2

[17] A. Gupta and L. Davis. Objects in action: An approach for combining action understanding and object perception. In *CVPR*, 2007. 1, 2

[18] A. Gupta, A. Kembhavi, and L. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE T. Pattern Anal. Mach. Intell.*, 31(10):1775–1789, 2009. 2

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

[20] H. Kjellstrom, J. Romero, and D. Kragic. Visual object-action recognition: Inferring object affordances from human demonstration. In *CVIU*, 2010. 1, 2

[21] M. Meila and J. Shi. Learning segmentation by random walks. In *NIPS*, 2000. 4

[22] K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the trees: a graphical model relating features, objects and scenes. In *NIPS*, 2003. 2

[23] L. Oakes and K. Madole. Function revisited: How infants construe functional features in their representation of objects. *Adv. Child Dev. Behav.*, 36:135–185, 2008. 1, 7

[24] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ALMC*, 1999. 4

[25] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *IEEE T. Pattern Anal. Mach. Intell.*, 34(3):601–614, 2012. 2

[26] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007. 2

[27] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3D human pose from 2D image landmarks. In *ECCV*, 2012. 3

[28] B. Russell, A. Efros, J. Sivic, W. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006. 2

[29] B. Scholkopf, A. Smola, and K.-R. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5):1299–1319, 1998. 4

[30] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures of parts. In *CVPR*, 2011. 2, 3, 5

[31] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010. 4

[32] B. Yao and L. Fei-Fei. Action recognition with exemplar based 2.5D graph matching. In *ECCV*, 2012. 3, 4

[33] B. Yao and L. Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE T. Pattern Anal. Mach. Intell.*, 34(9):1691–1703, 2012. 1, 2, 5