Zero-Shot Learning via Visual Abstraction

Stanislaw Antol[1], C. Lawrence Zitnick[2], and Devi Parikh[1]

[1] Virginia Tech, Blacksburg, VA, USA
[2] Microsoft Research, Redmond, WA, USA

Abstract. One of the main challenges in learning fine-grained visual categories is gathering training images. Recent work in Zero-Shot Learning (ZSL) circumvents this challenge by describing categories via attributes or text. However, not all visual concepts, *e.g.*, two people dancing, are easily amenable to such descriptions. In this paper, we propose a new modality for ZSL using *visual abstraction* to learn difficult-to-describe concepts. Specifically, we explore concepts related to people and their interactions with others. Our proposed modality allows one to provide training data by manipulating abstract visualizations, *e.g.*, one can illustrate interactions between two clipart people by manipulating each person's pose, expression, gaze, and gender. The feasibility of our approach is shown on a human pose dataset and a new dataset containing complex interactions between two people, where we outperform several baselines. To better match across the two domains, we learn an explicit mapping between the abstract and real worlds.

Keywords: zero-shot learning, visual abstraction, synthetic data, pose.

1 Introduction

Fine-grained object classification has gained significant attention in recent years. One of its main challenges is gathering training images. For example, though it may be easy to find images of birds, it might be very difficult to find images of specific species of birds, *e.g.*, "least auklet." Zero-Shot Learning (ZSL) [7,16,18, 30] addresses this scenario by providing an alternative approach that does not require any example training images. Instead, a user may provide other forms of side information, such as semantic visual attributes [16] (*e.g.*, "has black forehead," "has rounded wings") or textual descriptions of categories [7].

While semantic attributes or text-based descriptions provide an intuitive method for describing a variety of visual concepts, generating semantic descriptions is tedious or unreasonable for many visual concepts. For instance, how would a user semantically describe "a person sitting" to a recognition system that did not understand the concept of "sitting"? The problem is further exacerbated if the category is related to the interaction of multiple objects. For instance, consider the specific dancing poses between two people shown in the upper-right of Figure 1. Describing these scenes to a computer would require a lengthy textual description and still might not capture the full nuance.

D. Fleet et al. (Eds.): ECCV 2014, Part IV, LNCS 8692, pp. 401–416, 2014.

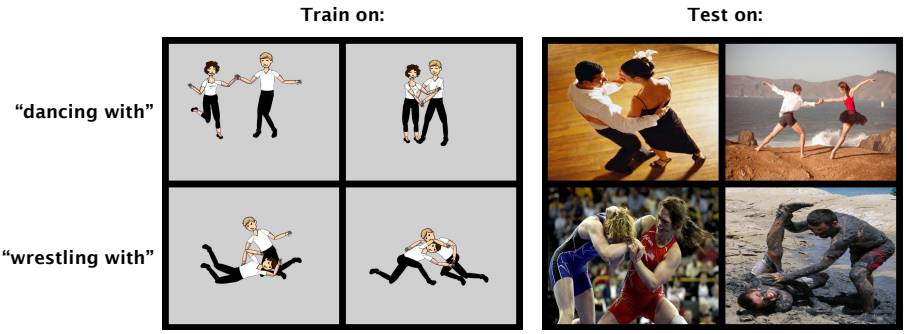

Fig. 1. Our approach: First, we have people create abstract illustrations (left) using our interface for various categories. Then we train models of these categories on illustration data. Finally, we test our models on real images (right).

To address this issue, we propose a new modality for ZSL that utilizes visual abstraction. The underlying intuition, shared by work in sketch-based image retrieval [6,24], is that it can be easier to communicate a visual concept through an abstract visual representation rather than a textual description. Thus, instead of a textual description like an attribute list, our proposed modality allows a supervisor to create abstract illustrations (left side of Figure 1).

In this paper, we use visual abstraction to train models for recognizing visual classes related to the pose of individuals or the interaction between two people. These concepts are of high interest in computer vision and are difficult to describe with traditional ZSL approaches. We introduce a novel image dataset, INTERACT, of 60 fine-grained interactions between pairs of people depicting various combinations of verbs and propositions (*e.g.*, "running after," "running to," "arguing with"). We test our approach on this dataset and a subset of the PARSE [21] dataset. We introduce a simple and intuitive interface that allows a supervisor to train these visual models. This interface lets users illustrate visual categories by varying the poses, expressions, gazes, and genders of people built from a collection of clipart. We present results for category-level ZSL (where each concept has a semantic name but is still hard to semantically *describe*, such as "dancing with") and instance-level ZSL (where each concept is very specific, such as a specific "dancing with" pose, and may not even have a semantic name). Surprisingly, our models, trained only on abstract illustrations (*i.e.*, visual abstractions), are effective at the category-level classification task on *real* images, even if only a single illustration depicting the concept is provided. As more example illustrations are provided, performance is further improved (Figure 4).

We create models that can generalize from abstract visualizations to real images by using a novel set of features. We analyze the role of different feature types (*e.g.*, contact between people, expressions) and show that some are more informative than others. When creating these example illustrations, users may visualize the semantically important aspects of the poses and interactions in more detail, but may have a fuzzier or even skewed notion of the other aspects.

Moreover, our easy-to-use interface results in biases in the illustrations (*e.g.*, the interface does not allow for out-of-plane rotation). To account for these human tendencies, as well as interface biases, we learn an explicit mapping from the features extracted from illustrations to the features extracted from real images. This allows us to improve performance on instance-level ZSL. Our visual abstraction interface, code, and datasets are publicly available.

2 Related Work

We discuss existing work on zero-shot learning, learning with synthetic data, learning semantic relations, pose estimation, and action recognition.

Zero-Shot Learning (ZSL): The problem of learning models of visual concepts without example images of the concepts is called Zero-Shot Learning. Attributes (mid-level, visual, and semantic features) [9,10,15,16] provide a natural interface for ZSL [16], where an unseen class is described by a list of attributes. Equipped with a set of pre-trained attribute classifiers, a test image can be probabilistically matched to each of these attribute signatures and be classified as the category with the highest probability. Instead of using a list of attributes, recent work [7] has leveraged more general textual descriptions of categories to build visual models of these categories. Our work takes a fundamentally different approach to ZSL. We propose a strictly visual modality to allow a supervisor to train a model for visual concepts that may not be easily *describable* in semantic terms, *e.g.*, poses of people, interactions between people.

Learning With Synthetic Data: Our work introduces the use of abstract visualizations as a modality to train visual models in a ZSL setting. Previously, papers have explored the use of synthetic data to aid in the training of vision algorithms. In many object recognition tasks, it is common to perturb the training data using affine warps to augment the training data [14]. Computer-generated scenes may also be used to evaluate recognition systems [13]. Shotton *et al.* [23] used synthetically generated depth data depicting humans to learn a human pose detector from this depth data. Unlike these approaches, we are trying to learn high-level, complex concepts where it is not feasible to automatically generate synthetic data, so we must rely on humans to create our synthetic data. Most similar to our work, the problem of semantic scene understanding using abstract scenes was studied in [31]. They use a dataset of simple sentences corresponding to abstract scenes to learn a mapping from sentences to abstract scenes. Recently, sequences of abstract scenes were used to predict which objects will move in the near future [11]. Unlike these works, we use abstraction to learn visual models that can be applied to *real* images. Sketch-based image retrieval [6, 24] allows users to search for an image by sketching the concept. Sketching complex interactions between people would be time consuming, and likely inaccurate for most lay users. More importantly, our modality has the potential to augment the abstract scenes with a large variety of visual cues (*e.g.*, gender, ethnicity, clothing, background) that would be cumbersome for users to convey via sketches.

Learning Semantic Relations: Previous papers have studied relations between people [26] and other objects [22, 29]. Most similar to the learning semantic relations part of our work, Yang *et al.* [26] used contact points to detect six different interactions between people. Sadeghi *et al.* [22] and Yao *et al.* [29] both model the relationship of people and objects when their combination creates a canonical pose or "visual phrase." Unlike all of this previous work, we are able to train our models for a larger number of concepts without relying on any training images. Several papers have studied the relations of people in groups using videos [4, 17], such as "queuing in line" or whether people are looking at each other [19]. While our approach only considers relations between people in the 2D image space, recently Chakraborty *et al.* [3] explored determining human relations using 3D information from a single image.

Pose Estimation and Action Recognition: Automatically estimating human pose [2, 27] and recognizing human actions [1, 28] in images has received a lot of attention in the vision community. These efforts are orthogonal to the focus of our work. We propose a new modality that enables us to train a vision system to recognize fine-grained interactions between people without any example images depicting those interactions. This can be augmented with any pose estimation technique at test time.

3 Datasets

To evaluate our approach, we need real images to *test* our visual models. For evaluation, we use two datasets: INTERACT, a new dataset that we introduce here, and the standard PARSE dataset [21].

3.1 INTERACT

Many fine-grained visual categories exist between *pairs of people*. While some datasets exist for a small number of these categories [26], we collected a new dataset with a significantly larger number of visual classes, INTERACT.

Interactions: Our dataset focuses on two people interacting via different verb phrases. They include transitive verbs (*e.g.*, "A is pushing B"), joint activities (*e.g.*, "A is dancing *with* B"), movement verbs with directional prepositions (*e.g.*, "A is walking *to* B"), and posture verbs with locational prepositions (*e.g.*, "A is sitting *next to* B"). We combine different verbs with different prepositions to get 60 verb phrases, including ones that share a verb but contain different prepositions, such as "running *to*" and "running *away from*." The full list of interactions can be found in the supplementary material (on the project website).

Real Image Collection: We crowdsourced our image collection on Amazon Mechanical Turk (AMT). We asked 3 workers to collect 20 images that meet the following criteria: they are all different photographs, they depict the sentence "Person A is *verb phrase* Person B," and all contain exactly 2 people. Note that we did not require them to have each person's entire body in the image

(*e.g.*, some body parts can be cropped out), which makes this dataset challenging (*e.g.*, right side of Figure 1). This resulted in 3,600 initial images.

Real Image Annotations: We also used AMT to collect various image annotations that are needed for our features via different custom interfaces. The pose annotation interface prompted the worker with one of our images and its corresponding sentence. We highlighted whether the worker should be annotating Person A or Person B in the sentence. The worker annotates the person's 14 body parts (right side of Figure 3). The worker provides their best guess if the part is occluded and responds "not present" if it is not within the image border. We had 5 workers annotate each person in each image and averaged them for the final ground truth pose annotations. In addition, workers annotated ground truth eye gaze (*i.e.*, looking to the image left or right), facial expression (*i.e.*, one of six prototypic emotional expressions [5] plus a neutral expression), and gender of each person via separate interfaces. We selected the mode of their responses for our final annotation. In addition to collecting the annotation of interest, two interfaces asked one additional question each. One asked if the prompted image contained exactly two main people or not and the other asked if the annotated pose overlaid on the prompted image was of good quality or not. We used the last two annotation queries to remove poor quality work. Additionally, a GIST-based [20] image matching scheme was used to remove duplicates. Removing these images gave us our final annotated dataset with 3,172 images (52.9 images per category on average). Some examples can be found in the bottom part of Figure 1 and the rightmost two columns of Figure 5. More details about our interfaces and our procedure can be found in the supplementary material.

3.2 PARSE

We also use a subset of the standard PARSE [21] dataset, which originally contains 305 images of *individuals* in various poses. We created a list of categories that frequently appear in the PARSE dataset (*e.g.*, "is dunking," "is diving for an object"). From the images that belong to these categories, we removed those that were used to train the pose detector [26]. Some categories (*e.g.*, "is standing") had disproportionately large number of images, so we removed images at random from these categories. This leaves us with 108 images in our dataset (7.7 images per category on average). We also collected the same annotations as in Section 3.1, except for pose (since ground truth pose annotations are already available with the dataset). See the supplementary material for more details.

4 Our Approach

In this section, we present our new modality for ZSL. We begin by introducing our user interface for collecting visual illustrations for training. We then describe the novel features that are extracted from our abstract illustrations and real images. Finally, we describe the approach used to train our models. The results of various experiments follow in Section 5.

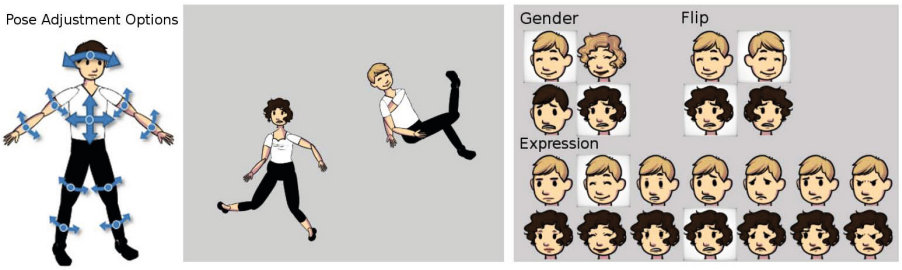

Fig. 2. User interface (with random initialization) used to collect abstract illustrations on AMT. Workers were able to manipulate pose, expression, gaze direction, and gender.

4.1 Visual Abstraction Interface

For our domain of interest, we conjectured that our concepts depend primarily on four main factors: pose, eye gaze, facial expression, and gender. Some other factors that we do not model, but may also be important are clothing, the presence of other objects, and scene context. A screenshot of our user interface is shown in Figure 2. Initially, two people (one blond-haired and one brown-haired) are shown with random poses, gaze directions (*i.e.*, "flip"), expressions, and genders. We allow our subjects to continuously manipulate the poses (*i.e.*, joint angles and positions) of both people by dragging on the various body parts. They may horizontally flip the people to change their perceived eye gaze direction. The facial expressions are chosen from the same selection as is used for the annotation of real images (Section 3.1). Finally, the subjects may select one of the two predominant genders for each clipart person.

To collect our training data for category-level ZSL, we prompt the user with a sentence to illustrate using the interface (*e.g.*, "Person A is dancing with Person B.", "A person is dunking."). To promote diversity, we encouraged them to imagine any objects or background, as long as the poses are consistent with the imagined scene (*e.g.*, a worker can imagine a chair and illustrate someone sitting on it). The interface includes buttons to annotate which clipart person corresponds to which person in the sentence. Some illustrations are shown on the left side of Figure 1 and in the left three columns of Figure 5. For the PARSE concepts, the interface is the same except that only one person is present.

For instance-level ZSL, we modify our previous interface. Instead of sentences, we first (briefly, for 2 seconds) show the user a real image and then they recreate it (from memory) as best they can. The stated goal is to recreate the real image so another person would be able to select the shown image from a collection of real images. This mimics the scenario when a person is searching for a specific image: they might be clear on the semantically important aspects while having a fuzzier or skewed notion of other aspects. Another bias of the illustrations occurs when it is impossible to recreate the real image exactly due to the limitations of the interface, such as not being able to change the height of the clipart people, the interface not allowing for out-of-plane rotation, *etc*.

4.2 Relation and Appearance Features

Using the annotations described in Section 3.1 (*i.e.*, pose, gaze, expression, and gender) for persons denoted by i and j, we compute a set of relation and appearance features. Some of our relation features are distance-based and some are angle-based. All distance-based features use Gaussians placed at different positions to capture relative distance. The Gaussians' σ parameters are proportional to the *scale* of each person. A person's scale is defined as the distance between their head and the center of their shoulders and hips. Unless otherwise noted, all angles/orientations are w.r.t. the image frame's x-axis. They are represented by 12 dimensional unit histograms with each bin corresponding to $\pi/6$ radians. Soft assignments are made to the histograms using linear weighting. The first two sets of features, Basic and Gaze, account for both people. The remaining five feature sets are described for a single person and must be evaluated twice (swapping i and j) and concatenated. The feature sets are described below.

Basic: This feature set encodes basic relation properties between two people, such as relative orientation and distance. We calculate each person's body angle (in the image frame). This is calculated from the image coordinates for the head and mid-point between shoulders. We place Gaussians at the center of the people and then use the distance between them to evaluate the Gaussian functions. We also calculate the angle (in the image frame) between the centers of the two people. This gives us a total of $2 * (12 + 1) + 12 = 38$ features. They can be thought of as simplifying the people into two boxes (possibly having different scale parameters) with certain orientations and looking at the relative positions and angle between their centers.

Gaze: The gaze feature set is encoded using 5 binary features, corresponding to i looking at j, j looking at i, both people are looking at each other, both people are looking away from each other, and both people are looking in the same direction. To determine if i is looking at j, we check if j's neck is in the *appropriate region* of the image. The image is divided into two parts by extending the line between i's head and neck and the appropriate region is defined to be the area where i is looking (which depends on i's gaze direction). Once we have both i looking at j and *vice versa* features, we compute the remaining three gaze features via the appropriate logic operations (*e.g.*, if i is looking at j and j is looking at i, then the looking-at-each-other feature is true).

Global: This feature set encodes the general position of the joints in reference to a body. Three Gaussians are placed in a 3×1 grid on the image based on the body's size and orientation (the blue circles in Figure 3). The positions of one person's 8 joints (two for each limb) are evaluated using all Gaussians from both Gaussian sets (*i.e.*, person i's joints relative to person i's global Gaussians and person j's global Gaussians), giving us a total of $8 * 3 * 2 = 48$ features.

Contact: This feature set encodes the specific location of the joints in reference to other body parts. For each person, we place Gaussians at 13 positions: 3

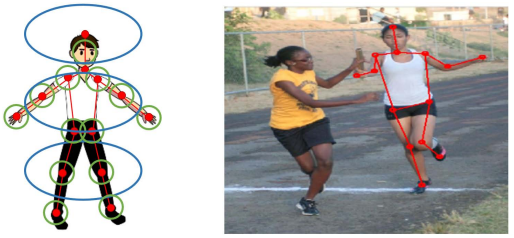

Fig. 3. (Left) Illustration of the part locations labeled for each person (red dots). We illustrate the standard deviations of the Gaussians used to compute normalized features (Section 4.2) for a joint's general location (blue circles) and whether there is contact with another part (green circles). (Right) The corresponding labeled part locations on a real image (from INTERACT) acquired from one of our annotation tasks.

for each limb and 1 for the head (the green circles in Figure 3). The positions of 8 joints (two for each limb) are then evaluated on the Gaussians placed on *themselves* (*e.g.*, is i's left hand near i's head) and the *other person* (*e.g.*, is i's right elbow near j's left shoulder) for a total of $8 * 13 * 2 = 208$ features.

Orientation: This final pose-based feature set encodes the relative (*i.e.*, *not* w.r.t. the image frame) joint angles. They are computed by finding the relative angle between parent and child joint positions (*e.g.*, left elbow and left shoulder). We compute 8 joint angles (two for each limb) for a total of $8 * 12 = 96$ features.

Expression: We convert expression into a set of 7 binary variables.

Gender: We convert gender into a set of 2 binary variables.

Thus, concatenating all of these features gives us a total of $38 + 5 + 2 * (48 + 208 + 96 + 7 + 2) = 765$ features, which are all between 0 and 1.

For the PARSE dataset, where each image only contains a single person, each of the previous feature sets are modified accordingly (if at all). The Basic feature set becomes only the body angle. The Gaze feature set is replaced with 2 binary variables that indicate if the person is looking to their left or right, respectively. The Global feature set is halved. The Contact feature set is also halved. This gives us a total of $12 + 2 + 24 + 104 + 96 + 7 + 2 = 247$ features.

By design, our illustration interfaces provide (without any additional annotation) the same data that was collected for the real images. Thus, we can use the same features for both abstract illustrations and real images.

4.3 Zero-Shot Learning Models

In category-level ZSL, we are trying to create a model that can classify an image as belonging to one of the given semantic classes (*e.g.*, "dancing with"). We use multiple one-vs-all linear Support Vector Machines (SVMs), trained on the abstract illustration features. At test time, these classifiers are used to determine the category of the real images. In instance-level ZSL, we are trying to decide

if an image represents a specific concept, *i.e.*, given a test real image, we wish to determine which specific abstract visualization (instance) corresponds to the real image. For this, we use Nearest Neighbor matching. Since our features are from two different domains, learning a mapping between them could improve the matching performance. This is described next.

4.4 Mapping From Abstract to Real for the Instance-level Model

We learn a mapping between the domain of abstract images and the domain of real images. To learn such a mapping, we need examples that correspond to the same thing in both domains. We use some of our instance-level illustrations (Section 4.1) as these abstract-real pairs. The mapping can learn to correct for both user and interface biases discussed in Section 4.1.

Simpler techniques, such as Canonical Correspondence Analysis [12], did not learn a good mapping between the abstract and real worlds. We found that General Regression Neural Networks (GRNN) [25] did better. We also found that converting from our abstract features into "real" features performed better than converting real features into "abstract" features. Thus, the GRNN's input is all of the abstract features and its output is all of the real features.

5 Experimental Results

In this section, we describe our experiments which show that our new modality for ZSL is able to create models that can learn category-level (Section 5.1) and instance-level (Section 5.2) visual concepts. We perform an ablation study on different feature sets, showing their performance contribution (Section 5.3). Finally, we utilize a state-of-the-art pose detector on both INTERACT and PARSE datasets to investigate our approach in a more automatic setting (Section 5.4).

5.1 Category-Level Zero-Shot Learning

We begin by experimenting with the ability of our novel modality to learn our category-level concepts, *i.e.*, classifying images into one of the semantic descriptions, such as "A is kicking B." To acquire the required training illustrations, we ran our visual abstraction interface with sentence prompts (described in Section 4.1) on AMT. We had 50 workers create an abstract illustration for each of the 60 semantic concepts from INTERACT (Section 3.1) and the 14 semantic concepts from PARSE (Section 3.2). After removing poor quality work, we are left with 3,000 and 696 illustrations, respectively.

The setup for all category-level ZSL experiments (unless otherwise noted) is described here. Using the abstract illustrations, we train multiple one-vs-all linear SVMs (liblinear [8]) with the cost parameter, C, set to 0.01, which worked reasonably well across all experiments. For INTERACT, there is ambiguity (at test time) as to which person is Person A and which person is Person B. To account for this, we evaluate each of the classifiers using both orderings, select the

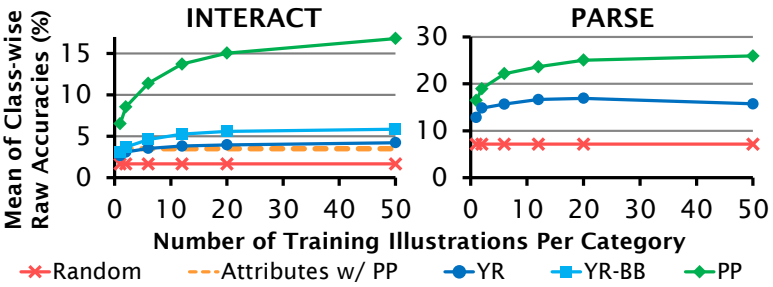

Fig. 4. We evaluate category-level ZSL performance on both datasets as described in Section 5.1. As we increase the number of ZSL training illustrations, our classification performance improves, but it begins to saturate. With both using perfect poses, our model (PP) does much better than the attribute DAP model (Attributes w/ PP). We also show results (described in Section 5.4) with output from a pose detector (YR) and, for INTERACT, a pose detector assisted by ground truth bounding boxes (YR-BB).

most confident score on orderings, and then predict the label of the classifier with the highest confidence. Our category-level classification metric is the mean of the class-wise raw accuracies. We observe performance as the number of training illustrations per category is increased. For each number of training illustrations, we average over 50 random selections of training illustrations (per category).

Results for our model with perfect poses, PP, are shown in Figure 4. It can be seen that even one illustration is able to perform several times better than random on both of our datasets. Adding additional training illustrations improves performance, although it begins to saturate around 20 training examples. For INTERACT, we reach ∼17% using all illustrations. We compare this to a stronger baseline: attribute-based ZSL. We define a vocabulary of 79 attributes, such as "A's hand is touching B's knee," gender, and gaze. For attribute ZSL, we use the DAP model from Lampert *et al.* [16]. Our approach significantly outperforms this approach (∼3.5%), demonstrating the benefit of our new modality for ZSL. More details about the baseline are in the supplementary material.

Some qualitative results are shown in Figure 5. Confusion matrices for the model are shown in the supplementary material. We also did a human agreement study on AMT for INTERACT. On average, the correct verb phrase for an image was selected only ∼51% of the time (averaged over 10 workers per image), which demonstrates how ambiguous this classification task can be. A similar human agreement study was done for the illustrations to identify the most canonical (*i.e.*, top) illustrations per category. Using the top illustrations instead of random ones to train our model provided modest improvements when using fewer training illustrations. If we treat any of the human labels that were collected during the INTERACT human agreement experiment as a valid label, we find that the PP model's performance increases to ∼37% (at 50 illustrations per category).

5.2 Instance-Level Zero-Shot Learning

We also test the ability of our new modality to learn instance-level concepts. To acquire the necessary training illustrations, we ran our visual abstraction interface with image prompts (as described in Section 4.1) on AMT. We showed a real image (one of 3,172 from INTERACT and one of 305 from PARSE) for two seconds to the workers, who recreated it using the interface. Through a pilot study, just as in [6], we found two seconds to be sufficient for people to capture the more salient aspects of the image. It is unlikely that a user would have every detail of the instance in mind when trying to train a model for a specific concept and we wanted to mimic this in our setup. We had 3 workers recreate each of the images, and after manually removing work from problematic workers, we are left with 8,916 and 914 illustrations for INTERACT and PARSE, respectively.

We perform classification via nearest-neighbor matching. If the real image's features match the features of any of the (up to) 3 illustration instances that workers created for it, we have found a correct label. We vary K, the number of nearest neighbors that are considered, and evaluate the percentage of real images that have a correct label within those K neighbors. We normalized K by the total number of illustrations. We need a training dataset to learn a mapping between the abstract and real worlds, *i.e.*, training the GRNN from Section 4.4. For INTERACT, we split the categories into 39 seen categories for training and 21 unseen categories for testing to minimize learning biases specific to specific categories (*i.e.*, verb phrases). The results are averaged over 10 random seen/unseen category splits. For PARSE, the training data corresponds to the 197 images

Fig. 5. The left columns show 5 random illustrations (of 50) used for classifier training. Columns 6 and 7 contain the most confident true positive and false positive for a given category, respectively. Mistakes include choosing a semantically reasonable verb (top), choosing the incorrect preposition (middle), and incorrect prediction due to the pose similarity between two classes (bottom). More examples are in the supplement.

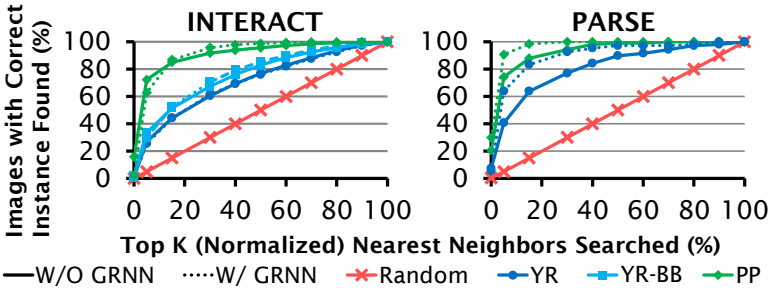

Fig. 6. We evaluate instance-level classification, showing the percent of images (y-axis) with the correct label being found within the top K (dataset size normalized) guesses (x-axis). We outperform random and see some benefit of learning a mapping (see Section 4.4) between the domains, particularly for the PARSE dataset.

that were not assigned a semantic category nor were used in the pose detector training (as discussed in Section 3.2). The training required for this mapping can be thought of as analogous to training the attribute predictors in the DAP model [16] for ZSL. One needs to train the attribute classifiers before one can use attribute descriptions with DAP. Similarly, we learn the mapping between the real and abstract world offline on a held out set of categories.

The results are shown in Figure 6. We see that our models are doing orders of magnitude better than chance just looking at the closest nearest neighbor and the gap increases as we search through neighbors that are further away. We also evaluate our approach by performing matching after transforming the train features via the GRNN (described in Section 4.4) with the GRNN's sole parameter, spread, set to 5 and 1 for INTERACT and PARSE, respectively. Using the mapping learned by the GRNNs helps, particularly on the PARSE dataset. More experimental results can be found in the supplementary material.

5.3 Feature Ablation Investigation

To better understand our system and the interactions in our dataset, we explore which of our features are most informative on INTERACT. Figure 7 shows the variation in performance when different feature sets are used. Comparing the first and third from the right bars, we note that our gaze features have negligible impact (actually performing slightly worse). This is possibly because, in real images, people might be looking head-on, whereas our abstract people can only look left or right. Of the appearance-based features, expression is most beneficial. This makes sense intuitively, since two people's poses can be roughly similar but the perceived action can change based on expression. For instance, when two people are wrestling *vs.* when they are hugging (*e.g.*, Figure 1). In both cases, arms can be around the other person's body, but expressions will change from angry to happy. Of the pose-based features, the Global feature is the most informative on its own. It indirectly captures contact and joint angles, so it is reasonable that it performs better than Contact or Orientation alone.

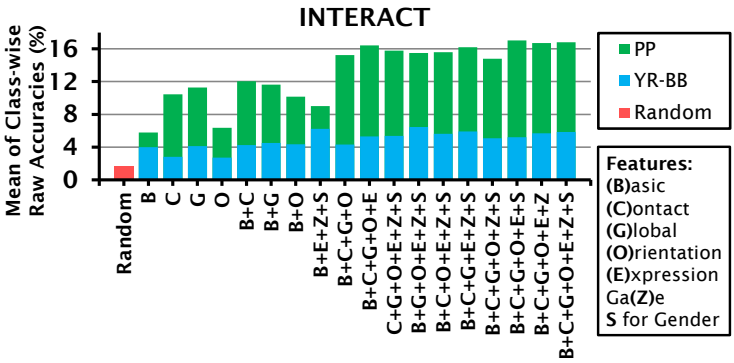

Fig. 7. We plot classification performance for INTERACT using different subsets of features. Some features, like Global, are more informative than others. Of the appearance-based features, Expression turns out to be most informative, presumably when body pose features are similar (*e.g.*, "wrestling" *vs.* "hugging").

5.4 Automatic Pose Evaluation

In this section, we do an evaluation of our category-level ZSL task using the current state-of-the-art pose detector developed by Yang and Ramanan [27]. We utilized the pre-trained PARSE model and detected the pose on both the INTERACT and the PARSE datasets. For the expression, gaze, and gender features, we continue to use human annotations. These results (YR) are shown in Figures 4, 6, and 7. As expected, due to the pose detector being developed for PARSE, automatic detection on the PARSE dataset yields reasonable performance (compared to perfect pose). The results on INTERACT do not perform nearly as well, although it still outperforms the baselines. To boost the performance of the pose detector on INTERACT, we also experimented with providing ground truth bounding boxes (YR-BB), which results in better performance.

INTERACT is significantly more challenging than PARSE for automatic pose detection. Thus, it is not surprising that incorrectly detected poses confuse our models. Properties that make INTERACT particularly challenging include: images from arbitrary perspectives, more difficult (for the detector) poses (*e.g.*, "crawling," "lying"), overlapping people (*e.g.*, "hugging," "standing in front of"), and incomplete poses (*i.e.*, not all body parts are present). We investigated this latter point by selecting images from INTERACT based on the number of parts present in the image. There are 14 parts per person and we ensure that both people have at least a certain number of parts. Requiring all parts to be within the image reduces INTERACT to 1,689 images (from 3,172). 91.5% of our images contain at least 7 parts per person. More of these details can be found in the supplementary material. We re-evaluate our category-level ZSL performance (at 50 training illustrations per category) as we vary the part threshold and show our results in Figure 8. Although there is some noise, both the perfect pose and automatic pose detection methods show an increase in accuracy as we require

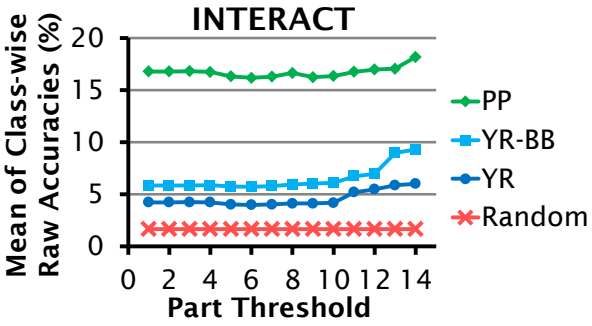

Fig. 8. We plot classification performance as we vary the minimum number of parts both people have (see Section 5.4). While the perfect pose approach only has minor improvements, both detection approaches improve more when all people are fully visible, showing that current detectors do not work well when parts are missing.

more parts to be within the image border. This result suggests that there is a lot to gain by furthering research into more robust pose detectors. We would like to reiterate that we have part annotations even in the case of occlusion, which probably accounts for some of the automatic detector's performance difference with perfect poses. We hope the introduction of INTERACT will help the community advance pose detectors in more practical settings.

6 Conclusions

We propose a new modality for Zero-Shot Learning (ZSL) of concepts that are too difficult to describe in terms of attributes or text in general (*e.g.*, "holding hands with"). A user illustrates the concept through visual abstraction. We demonstrate its utility for classifying poses of people and interactions between two people. We introduce a new dataset containing 60 fine-grained verb phrases describing interactions between pairs of people. We present results for category-level ZSL (where the concept has a semantic name, *e.g.*, "crouching with") and instance-level ZSL (where the concept is specific and may not have a semantic name). We report results on our new dataset, as well as a standard single-person pose dataset. We also learn a mapping from abstract-world features to real-world features. Our approach outperforms several baselines. We also analyze the information captured by various subsets of features, such as contact and expression. Our interface, code, and datasets are made publicly available.

Acknowledgements. We thank Yi Yang and Deva Ramanan for making their implementation publicly available. We thank Micah Hodosh and Julia Hockenmaier for their initial discussions. We also would like to thank the hundreds of Turkers, without whom this work would not have been possible.

References

1. Ali, S., Shah, M.: Human action recognition in videos using kinematic features and multiple instance learning. PAMI (2010)
2. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: ICCV (2009)
3. Chakraborty, I., Cheng, H., Javed, O.: 3d visual proxemics: Recognizing human interactions in 3d from a single image. In: CVPR (2013)
4. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: CVPR (2011)
5. Darwin, C.: The Expression of the Emotions in Man and Animals. Oxford University Press (1998)
6. Eitz, M., Hildebrand, K., Boubekeur, T., Alexa, M.: Sketch-based image retrieval: Benchmark and bag-of-features descriptors. IEEE Transactions on Visualization and Computer Graphics (2011)
7. Elhoseiny, M., Saleh, B., Elgammal, A.: Write a classifier: Zero-shot learning using purely textual descriptions. In: ICCV (2013)
8. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. JMLR (2008)
9. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: ICCV (2009)
10. Ferrari, V., Zisserman, A.: Learning visual attributes. In: NIPS (2007)
11. Fouhey, D.F., Zitnick, C.L.: Predicting object dynamics in scenes. In: CVPR (2014)
12. Hotelling, H.: Relations between two sets of variates. Biometrika (1936)
13. Kaneva, B., Torralba, A., Freeman, W.T.: Evaluation of image features using a photorealistic virtual world. In: ICCV (2011)
14. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
15. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV (2009)
16. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
17. Lan, T., Wang, Y., Yang, W., Mori, G.: Beyond actions: Discriminative models for contextual group activities. In: NIPS (2010)
18. Larochelle, H., Erhan, D., Bengio, Y.: Zero-data learning of new tasks. In: AAAI (2008)
19. Marin-Jimenez, M., Zisserman, A., Eichner, M., Ferrari, V.: Detecting people looking at each other in videos. IJCV (2013)
20. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV (2001)
21. Ramanan, D.: Learning to parse images of articulated bodies. In: NIPS (2007)
22. Sadeghi, M.A., Farhadi, A.: Recognition using visual phrases. In: CVPR (2011)
23. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR (2011)
24. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. PAMI (2000)
25. Specht, D.F.: The general regression neural network-rediscovered. Neural Networks (1993)

26. Yang, Y., Baker, S., Kannan, A., Ramanan, D.: Recognizing proxemics in personal photos. In: CVPR (2012)
27. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR (2011)
28. Yao, B., Fei-Fei, L.: Action recognition with exemplar based 2.5d graph matching. In: ECCV (2012)
29. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR (2010)
30. Yu, X., Aloimonos, Y.: Attribute-based transfer learning for object categorization with zero/one training example. In: ECCV (2010)
31. Zitnick, C.L., Parikh, D., Vanderwende, L.: Learning the visual interpretation of sentences. In: ICCV (2013)