Recognition Using Visual Phrases

Mohammad Amin Sadeghi[1,2], Ali Farhadi[1]
[1]Computer Science Department, University of Illinois at Urbana-Champaign
[2]Computer Vision Group, Institute for Research in Fundamental Sciences (IPM)
{msadegh2,afarhad2}@illinois.edu

Abstract

In this paper we introduce visual phrases, complex visual composites like "a person riding a horse". Visual phrases often display significantly reduced visual complexity compared to their component objects, because the appearance of those objects can change profoundly when they participate in relations. We introduce a dataset suitable for phrasal recognition that uses familiar PASCAL object categories, and demonstrate significant experimental gains resulting from exploiting visual phrases.

We show that a visual phrase detector significantly outperforms a baseline which detects component objects and reasons about relations, even though visual phrase training sets tend to be smaller than those for objects. We argue that any multi-class detection system must decode detector outputs to produce final results; this is usually done with non-maximum suppression. We describe a novel decoding procedure that can account accurately for local context without solving difficult inference problems. We show this decoding procedure outperforms the state of the art. Finally, we show that decoding a combination of phrasal and object detectors produces real improvements in detector results.

1. Introduction

How should one detect complex visual composites, for example "a person riding a horse"? Conventional wisdom suggests detecting components like "person" and "horse" independently, and then describing the relation. This approach is motivated by the very large number of composites that can be built by very few basic atoms. Also, there will be very few training examples for most composites due to the increase in specifications.

The main weakness of this argument is that the appearance of the objects may profoundly change when they participate in relations. For example, people riding horses take relatively few postures, as do horses with people on their back. Relations may also create important occlusion regularities. For instance, one leg of the rider is often occluded by the horse. As a result, visual composites might be much easier to detect than their participant components. One extreme example is a scene (e.g. kitchen). There are quite good "kitchen" classifiers, but none proceeds by finding "toaster", "coffeepot", and "kettle", then fusing.

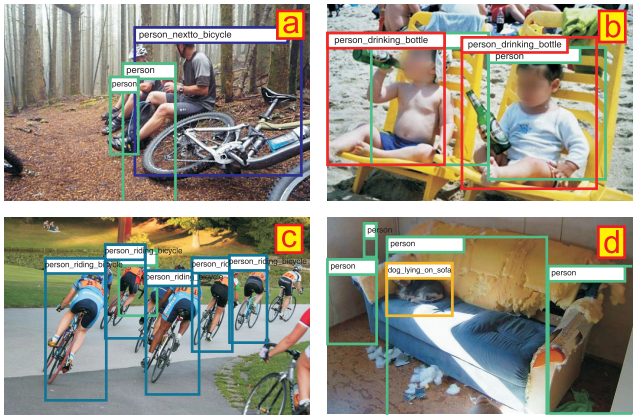

Figure 1. Detecting visual phrases is often significantly more accurate than detecting participating objects. In image "a", the bicycle detector and the person detector do not have accurate responses whereas our "person next to bicycle" detector correctly finds the visual phrase. In image "b", the bottle detector does not produce any sensible detection while our "person drinking from bottle" detector accurately finds instances of the visual phrase. The faces of the children are blurred here due to privacy concerns. In image "c", the person detector could only find one instance of a person while our "person riding bicycle" detector finds 5 instances correctly. In image "d", neither the dog detector nor the sofa detector are producing reliable responses but our "dog lying on sofa" detector finds the visual phrase correctly. We believe that detecting visual phrases are often much easier than the participating objects as visual phrases exhibit less visual complexity. See Figure 4, and Table 1 for quantitative evaluations.

Surprisingly, in the literature, there is no composite intermediate between objects and scenes. In this paper, we introduce such intermediate composites, which we call "***visual phrases***". Visual phrases correspond to chunks of meaning bigger than objects and smaller than scenes. We show that the reduction in the visual complexity exhibited by visual phrases is often so great that very accurate detectors can be trained with little training data. For example, our "person

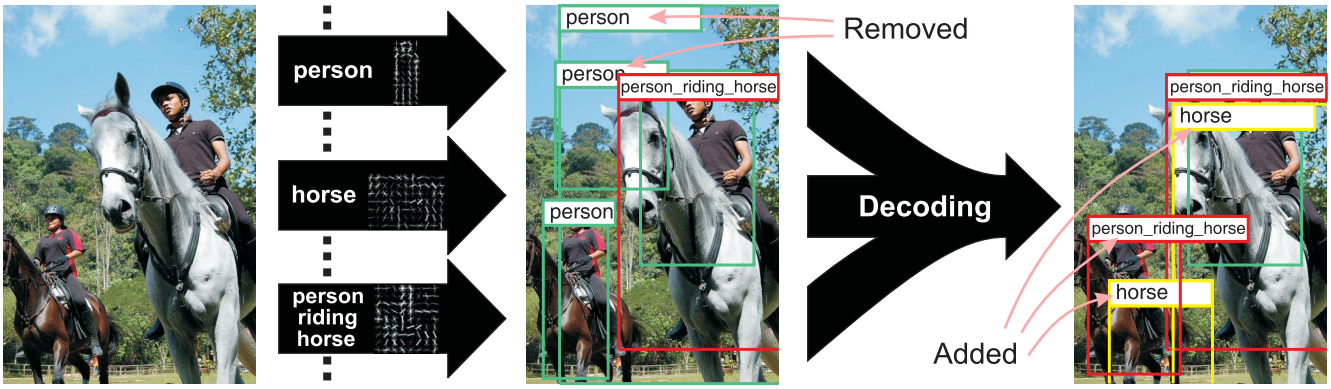

Figure 2. We use visual phrase and object models to make independent predictions. We then combine the predictions by a decoding algorithm that takes all detection responses and decides on the final outcome. Note that a) Visual phrase recognition works better than recognizing the participating objects. For example, the horse detector does not produce reliable predictions about horses in this picture while the "person riding horse" detector finds one instance; b) Our decoding then successfully adds two examples of horses and removes two wrong predictions of people by looking at other detections in the vicinity.

riding horse" detector works much better than "person" and "horse" detectors while using less training data (see Figure 4 for experimental data). Figure 1 shows examples of the cases where best object detectors miss objects while the visual phrase detectors correctly localize visual phrases.

One reasonable concern is that the number of phrases grows exponentially in the number of objects, and there may not be enough training data for each visual phrase. Our experience of visual phrases mirrors the experience of machine translation community with linguistic phrases. The number of useful visual phrases (phrases) is significantly smaller than the number of all possible combinations of objects (words). There are many visual phrases that could occur during tasks but we tend to encounter very few of those. Further, many visual phrases show substantially reduced visual complexity compared to independent objects and so one doesn't need to have a large number of training examples to accurately learn visual phrases. For example, our "person riding horse" detector, learned with default settings on only 50 positive examples, significantly outperforms the heavily fine tuned state of the art models for "horse" and "person" learned on thousands of examples (see Figure 4 and Table 1 for more details).

We believe that the current choice of categories as basic atoms of recognition is arbitrary. We argue that these basic atoms should be chosen by performance criteria. Opportunism is the key to this principle. Instead of learning some basic level detectors and using them no matter how good they are, we learn detectors at different levels and use reliable ones and then decode to obtain a final interpretation (Figure 2). Decoding uses all detection responses to decide which detections are worth reporting as the final result. Decoding is an inevitable part of multiple object detection. The decoder may need to boost some detections and suppress others based on local context.

There is an analogy to machine translation problems where the alignment has to be established between phrases and areas of images. One might think of our system as having a phrase table with entities like "person", "horse", and "person riding horse". The ultimate goal is to look at all phrases and find the longest phrase that matches. This procedure is often called decoding in machine translation. Our decoder has to take into account that some of the detectors should overlap and when they overlap it has to decide which of the overlapping detectors are worth reporting.

In this paper we show the benefits of opportunistically selecting basic atoms of recognition and the significant gain in directly detecting visual phrases. Our contributions are: 1) Introducing visual phrases as categories for recognition; 2) Introducing a novel dataset for phrasal recognition; 3) Showing that considering visual phrases provides a significant gain over state of the art object detectors coupled with the state of the art methods of modeling interactions; 4) Introducing a decoding algorithm that takes into account specific properties of interacting objects in multiple levels of abstraction; 5) Producing state of the art performance results in multi-class object recognition.

2. Related Works

Object Recognition: Due to limited space we only mention the most relevant works in object recognition. Deformable templates [3, 4] and part based models [1, 10, 5] are of the most successful methods in object recognition. In this paper we use the state of the art detectors in [9] using deformable part models. This work considers multiple roots to model the appearance changes due to viewpoint or inherent intra-class variations.

Object Interactions: All methods that model interactions between objects neglect the change in the appearance of objects due to interactions with other objects. We differ from all by taking this effect into account. Gupta et. al. [11] model these interactions by modeling the prepositions and adjectives that relate nouns. Yao and Li [16] model the

interactions between human pose and objects by coupling the human pose estimation and object recognition together. In [7] the interactions between objects is modeled implicitly in the context of predicting sentences for images. The most relevant to ours is the work by Desai, Ramanan, and Fowlkes[2]. They encode the interactions between objects by a set of relationships like "on the right of", "on the left of", "on the top of", etc. They then learn a weight for the interactions of objects in each of these relationship bins and use them to re-weight the confidence of detectors. We differ from them as we consider the change in appearance of interacting objects. We show that neglecting the change in the appearance of interacting objects causes recognition issues, while modeling it significantly improves recognition results.

Scene Understanding has been one of the mainstream tasks in computer vision. One natural approach is to represent scenes as with global features that take into account general information about images [15, 13]. An alternative is to consider objects in the scene and discover clusters of correlated objects [14]. Objects in scenes are not independent and tend to cluster. We think these clusters might be formed at the phrase level as well. There is a neglected semantic gap between scenes and objects. We introduce visual phrases to cover this gap.

Machine Translation aims at automatic translation from one language to another one. Statistical translation methods are among successful approaches. In the common architecture of statistical translation models, there is a translation model, a language model, and a decoding algorithm. The decoding algorithm has to decide the final translation given the translation model, language model, and a query sentence. Word based translations are usually not desirable as there is no direct mapping between words across languages and syntactic differences are significant. However, phrasal translations, which are the inspirations of this work, are fashionable in machine translation because they allow multiple to multiple translations, use local context in translation, and allow translation of non-compositional phrases[12].

3. Phrasal Recognition

Our task is to learn appearance models not only for basic level categories but also for richer levels of abstractions, visual phrases. Having learned these appearance models, we show significant gains in considering some visual phrases as a whole instead of detecting the basic atoms and then modeling the interactions, see Figure 4 and Table 1. We also consider the problem of object recognition in a multi-class framework and model the interactions between categories which includes objects and visual phrases. We show significant boost in multi-class recognition performance using our decoding method along with our visual phrase models comparing to the state of the art basic level models coupled with the state of the art interaction models.

Figure 3. The phrasal recognition dataset consists of 17 phrases and 8 objects. There are 2769 images in this dataset and on average 120 images per category. This figure shows 6 example of 7 different visual phrases in our dataset. Rows correspond to visual phrases: dog jumping; horse and rider jumping; person drinking from bottle; person jumping; person lying on beach; person lying on sofa; person next to bicycle.

To this end, we need to have a dataset of phrases and objects. There are multiple datasets available for object recognition. Unfortunately, there is no test bed suitable for phrasal recognition. Here, we introduce the first phrasal recognition dataset.

3.1. Phrasal Recognition Dataset

We first select 8 object classes from Pascal VOC2008 dataset [6] that are suitable for modeling the interactions between objects: person, bike, car, dog, horse, bottle, sofa, and chair. We then add a list of 17 visual phrases using 8 selected object classes. Our visual phrases are formed by either an interaction between objects or activities of single objects. These visual phrases are: person riding horse; person sitting on sofa; person sitting on chair; person lying on sofa; person lying on beach; person riding bicycle; horse and rider jumping; person next to horse; person next to bicycle; bicycle next to car; person jumping; person next to car; dog lying on sofa; dog running; dog jumping; person running; and person drinking from a bottle. We also add a background class.

We use Bing image search to gather images for the phrases and manually filter out irrelevant images. For basic level categories we used Pascal images. We manually obtain bounding boxes of all the 8 objects along with 17 phrases for all of the images in the dataset. There are 2769 images (822 negative images) in our dataset and on average each class has 120 examples. In total there are 5067 bounding boxes (1796 for visual phrases+3271 for

objects) in this dataset. As expected, the number of training examples decreases as the complexity of the phrase increases. However, the collapse in the visual complexity of phrases is so great that one doesn't need to have many training examples to learn visual phrases(see section 3.2). This dataset and the phrase models are publicly available at http://vision.cs.uiuc.edu/phrasal/. Figure 3 shows examples of images in our dataset.

3.2. Appearance models

The appearance models for each category, including objects and visual phrases, are learnt using the latest version of deformable part models [8]. We learn these models for each of our 17 phrases in our dataset using provided bounding boxes. Available models on the 8 categories from Pascal [8] are used as models for objects in the phrasal recognition dataset. We use these models to evaluate the benefits of phrasal recognition. Many of visual phrase detectors have accurately learned the phrase, Figure 4. This is mainly due to the fact that often the appearance of visual phrases has limited variance comparing to the objects in the phrase. For the same reason, the number of necessary training examples for training appearance models for visual phrases can often be very small. Similar to object detectors, some of the visual phrases are hard to train as they have higher variance in the appearance.

4. Decoding Multiple Detections

Decoding takes all detector responses as input and decides on the final outcome. Non-maximum suppression (NMS) is the usual form of decoding. Perfect detectors with excellent tightly tuned models should seldom, if ever, need decoding because there is no ambiguity in what to report. Current detectors are not perfect so decoding is a necessary part of every multiclass object detection method.

One natural decoding strategy, which outperforms NMS, is to model the interaction between objects by having pairwise terms in the scoring function [2]. This approach often yields intractable inferences and one needs to greedily search the space of labels. Pairwise terms are used to model interactions between objects resulting in fiercely intractable combinatorial problems which are hard to approximate.

Our philosophy is that well designed feature representations should make it unnecessary to account for pairwise interactions. To do that, detector responses should be aware of other detectors in a vicinity. We explicitly encode this in our feature representation resulting in very fast, exact inference methods.

Notation: Following the notation of [2], an image is represented as a collection of overlapping bounding boxes which are represented by features x_i. Write $X = \{x_i : i = 1...M\}$ as the representation of an image where M is the total number of bounding boxes for an image. To get these bounding boxes we run all of the detectors on all of the images. For each bounding box, we know its position, scale, and the confidence of the detector that reported this bounding box. We also assume that there are K different categories and $y_i \in \{0, 1\}$ is the label for each bounding box. $y_i = 1$ means that the i^{th} bounding box should be considered in the final response and $y_i = 0$ is otherwise. $Y = \{y_i : i = 1...M\}$ is the entire label for image X. $c_i \in \{0, 1, ..., K\}$ is the indicator variable showing the category detector that selected the i^{th} bounding box. The score of labeling image X with label Y is defined as $S(X, Y) = \sum_i w_{c_i}^T x_i$ where i is the index to the i^{th} bounding box in image X and w_{c_i} is the set of weights that corresponds to the class of the i^{th} bounding box. We do not consider the pairwise relationships in the scoring function as these relationships are encoded in our feature representation (section 4.1).

4.1. Representation

We expect our final score for each bounding box to be aware of the results of all other categories nearby. We explicitly encode this in our feature representations. Our representation of an image is based on representations of bounding boxes obtained on each image using all detectors and consists of confidences, the amount of overlap and size ratio of neighboring bounding boxes. To do that we run all of our detectors on each of the images. We consider three spatial relationships: above, below, and overlapping. For each window, for each category, and for each of these spatial bins we consider the confidence of the best scoring window, its overlap, and its size ratio to the represented window. We also add the confidence of the represented window to the features. This means that our representation has $K \times 3 \times 3 + 1$ dimensions.

4.2. Inference

We assume bounding boxes are independent given their features. Our feature design makes this assumption reasonable and so our inference is exact. Our inference is

$$\begin{aligned} Y^* &= \{y_i^*, i = 1...M\} \\ {y_i}^* &= \arg\max_{y_i} w_{c_i}^T \Phi(X, y_i) \end{aligned} \quad (1)$$

where i is the index to bounding boxes and w_{c_i} is the corresponding weights for the class of the i^{th} bounding box and $\Phi(X, y_i)$ generates features for that bounding box. This is very simple exact inference as $y_i \in \{0, 1\}$ and y_i's are independent.

4.3. Learning

Our model is a form of max margin structure learning. The structured label Y has to be predicted using our decoding model. The objective function takes the form of:

$$\min_{w,\xi} \sum_{c \in \{0,...,K\}} \frac{1}{2} \parallel w_c \parallel_2^2 + \lambda \sum_n^N \xi_n \quad (2)$$

$$s.t. \forall n, H_n,\ S(X_n, Y_n) - S(X_n, H_n) \geq L(Y_n, H_n) - \xi_n$$

where $n \in \{1, ..., N\}$ is the index to the image and L is the loss between the hypothesis $H_n = \{h_{n,i}, h_{n,i} \in \{0,1\}, i = 1...M\}$ and the true structured label Y_n, ξ_n is a slack variable, and λ is the tradeoff between the regularization and loss. This max margin formulation requires all of the hypotheses to score lower than the ground truth labels by at least the amount of loss. We model the loss as hamming loss. Eq. 2 can be reformulated as

$$\min_{w} \sum_{c \in \{0,...,K\}} \frac{1}{2} \parallel w_c \parallel_2^2 + \tag{3}$$

$$\lambda \sum_{n}^{N} \sum_{i}^{M} w_{c_i}^T (\phi(X_n, h_{n,i}^*) - \phi(X_n, y_{n,i})) + L(H_n^*, Y_n)$$

$$\text{s.t. } H_n^* = \arg\max_{H_n} \sum_{i}^{M} w_{c_i}^T \phi(X_n, h_{n,i}) + L(H_n, Y_n) \tag{4}$$

Fortunately, in this min-max formulation, our inner maximization is exact and very fast. We solve this optimization problem by subgradient descent method as follows.

We first randomly initialize w_{c_i}'s and solve for H^*'s in the inner maximization problem, Eq 4. This is an easy maximization as $h_i \in \{0,1\}$ and the labels for bounding boxes are independent given their features. We then fix the H^*'s and use the subgradient of the objective function to minimize it. The step size is $1/t$ where t is the number of iterations. Having taken one step, we fix w_{c_i}'s and search for H^* again. We iterate till we converge. The convergence criteria is set by looking at the consecutive improvements on the objective value.

When converged, we use $w_{c_i}^*$ in the inference model (Eq. 1) to rescore the bounding boxes accordingly and also infer the final labels.

5. Results

To evaluate phrasal recognition and our decoding method we show extensive quantitative results on two tasks: a) single category detection, and b) decoding: multi-category detections. We compare our results to state-of-the-art performance results in both tasks.

5.1. Single Category Detection

We use deformable part models with default settings [9] to train detectors for our 17 visual phrases. For objects we use the trained models from [8]. These models produce state of the art results in the single object detection task on Pascal dataset. We show significant gain in modeling the visual phrases comparing to separately detecting participating objects and then modeling the relations. Figure 4 shows Precision-Recall (PR) curves for some of the visual phrase detectors. We trained these detectors with at most 50 positive examples. Many of the visual phrase detectors produce promising results. To further demonstrate the substantial gain in considering visual phrases, we compare our visual phrase detectors with a baseline that tries to best model interactions between objects.

Phrases (Trained with 50 positive images)	Phrase (AP)	Baseline (AP)	Gain (AP)
Person next to bicycle	0.466	0.252	**0.214**
Person lying on sofa	0.249	0.022	**0.227**
Horse and rider jumping	0.870	0.035	**0.835**
Person drinking from bottle	0.279	0.010	**0.269**
Person sitting on sofa	0.262	0.033	**0.229**
Person riding horse	0.787	0.262	**0.525**
Person riding bicycle	0.669	0.188	**0.481**
Person next to car	0.443	0.340	**0.103**
Dog lying on sofa	0.235	0.069	**0.166**
Bicycle next to car	0.448	0.461	**-0.013**
Dog Jumping	0.072	0.134	**-0.062**
Person sitting on chair	0.201	0.141	**0.060**
Person running	0.718	0.484	**0.234**
Person lying on beach	0.179	0.140	**0.039**
Person jumping	0.317	0.036	**0.281**
Person next to horse	0.351	0.287	**0.064**
Dog running	0.504	0.160	**0.344**

Table 1. AP scores for all of the visual phrases in our dataset. We compare our visual phrase detection results with a baseline detector that consists of the state of the art object detectors coupled with an operator that tries to best model the relationships between objects. This baseline is biased toward the best possible outcome on the test set. Please see section 5.1 for more details on the baseline. Note the significant gain (third column) in using visual phrases compared to an optimistic upper bound for detecting objects and modeling their relations. Some of the visual phrase detectors like "horse and rider jumping", "person riding horse", "person riding bicycle" show amazing gain. At the same time, some of the visual phrase detectors like "bicycle next to car" doesn't work as well. We demonstrate an opportunistic principle for selecting what detectors to use based on performance. See section 4.

The baseline takes the confidence responses of participating object detectors as input and tries to best model the interactions between the objects. It is challenging to build a perfect detector that takes into account interactions of objects. We, therefore, build a baseline detector that performs on the *test set* as best as it can. The performance of the baseline can be regarded as an optimistic upper bound on how well one could detect visual phrases by detecting participating objects using the best current detectors. We run detectors for each of the participating objects and consider overlapping responses. There are multiple ways of modeling the interactions between objects: a) We extend the bounding boxes of the overlapping responses of participating objects to estimate the bounding box of the visual phrase. We then compute the average of the confidences of the bounding boxes of the participating objects to estimate a score for the estimated bounding box. We then use this score to produce the PR curves. b) This is similar to "a" but we consider the minimum of the confidences of partic-

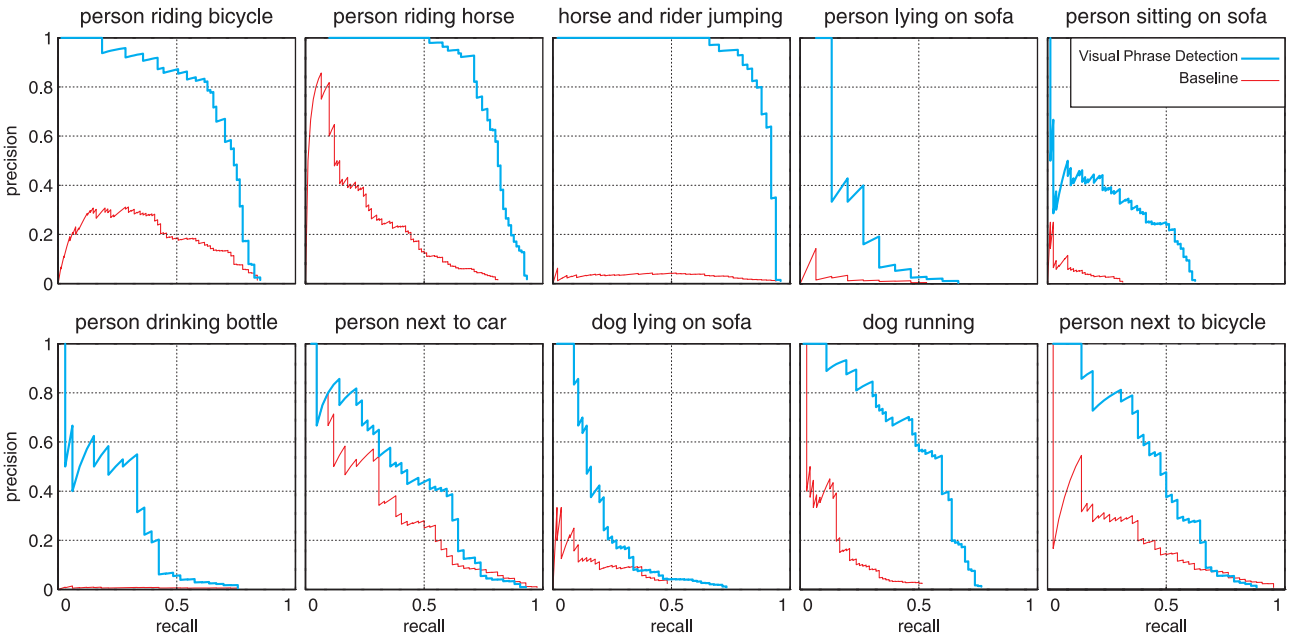

Figure 4. Precision-Recall curves for detecting 10 visual phrases in our dataset comparing to the baseline. The comparison to this baseline is biased toward best possible outcome on the test set. Please see section 5.1 for more information on the baseline. Note the significant gain in detecting visual phrases compared to detecting objects and describing their relations. The gain is astonishing because the phrase detectors are trained using at most 50 positive training examples with default settings while the object detectors are heavily fine tuned and trained using thousands of examples. Further, The baseline is heavily biased toward best possible outcome on the test set. Please see Table 1 for detailed AP's for all of the visual phrases in our dataset.

ipating objects rather than their average. c) This is similar to "a" and "b" but we use maximum confidence instead of the average or the minimum. d) We regress the position, scale, and confidence of the final phrase prediction against the positions, scales, and confidences of the participating objects on the *test set*. To produce the best possible outcome, we run all of these procedures and pick the one that best performs on the *test set*. Estimates of performance of this baseline are generous because we choose a combination that best performs on the *test set*. To be more conservative, we run the baseline with two sets of detectors (state-of-the-art models in [9] trained on our dataset, and state-of-the-art models in [8]) and pick the best one.

To evaluate our phrase detectors we test each of the visual phrase models and the corresponding baseline detector on a test set of approximately 200 images. Each test set has roughly 50 positive and 150 negative examples. The negative images are selected in a way that they do not contain any example of participating objects. For phrases that have only one participating object the baseline would be the corresponding models from [8].

Figure 4 depicts comparisons between the visual phrase detection results and the baseline. Note the significant improvements using visual phrase detectors trained on only 50 positive examples and default settings compared to heavily fine tuned object detectors [8] trained on thousands of examples. Further, the baseline is learned on the test set. Table 1 shows Average Precision (AP) for all of the visual phrase detectors compared with the results of the baseline detectors. In most cases our visual phrase detectors are outperforming the baseline detectors by significant margins despite the fact that the baseline is designed to perform best on the test set. There are visual phrases like "dog jumping" where neither the visual phrase detectors, nor the baseline detectors have promising results. These are hard objects and visual phrases with unmanageable variance in appearance. The results in Figure 4 and Table 1 support of the neglected fact that the appearance of the objects may change when they interact. Figure 4 and Table 1 show amazing gains when considering visual phrases.

5.2. Decoding

We compare our decoding algorithm with that of [2] on our phrase dataset. This is to evaluate our decoding method with other decoding methods not to evaluate the merits of phrasal recognition as all of our detectors, including visual phrase detectors, are provided as input to all decoding methods. We run all of the detectors for all of the phrases as well as the objects and construct the features as explained in section 4.1. We then use our decoding algorithm to learn a set of weights that rescore the confidences of the bounding boxes based on interactions. We compute per class AP, overall AP and mean per image AP for comparisons. We also learn the model of [2] using the publicly available code on our dataset. We again rescore the confidences of the bounding boxes using the weights provided by this model

and compute per class AP, overall AP and mean per image AP. All these three decoding procedures are learned on visual phrases as well as objects. Our decoding gets an overall AP of 0.319 and mean per class AP of 0.495 comparing to the overall AP of 0.313 and mean per class AP of 0.493 for [2] and AP of 0.308 and mean per class AP of 0.491 for NMS using models in [9]. We believe that encoding the interactions in the representation makes the models more manageable comparing to encoding the interactions by pairwise terms in the model and so resulting in better performance in decoding.

5.3. Phrasal Recognition Helps Object Detection

We learn our decoding and the method of [2] using only the objects (not phrases) and compare it with the case when we consider both phrases and objects. Table 2 shows per class AP's for both our decoding and that of [2] with and without phrases. Significant gains in the performance of detectors when coupled with visual phrases establish the importance of visual phrases coupled with reliable decoding.

Our decoding helps recognition of single objects using phrases. For example, in image "a" of Figure 6, a confident "person riding bicycle" detector helps boosting the bicycle detection and suppressing wrong person predictions. Object detections also help visual phrase recognition. For example, in image "b" of Figure 6, the confident sofa detector boosts the confidence of the "dog lying on sofa" detections.

6. Discussion

In this paper, we introduce visual phrases, show significant gains in considering them, introduce the phrasal recognition dataset, and a decoding algorithm that outperforms state of the art methods. Building long enough phrase tables is still a challenge.

The dimensionality of our features grows with the number of categories. However, there is no need to consider all of the categories when we model the interactions. For this reason, one might only consider a fixed number of categories for each bounding box.

We speculate that the relations between attributes and objects, parts and objects, visual phrases and scenes, and objects and visual phrases mirror one another. Future work will investigate systems to decode complete sets of detections covering the semantic spectrum.

7. Acknowledgments

This work was supported in part by the National Science Foundation under IIS -0803603 and in part by the Office of Naval Research under N00014-01-1-0890 and under N00014-10-1-0934 as part of the MURI program. Ali Farhadi was supported by Google Ph.D fellowship. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of NSF, ONR, or Google.

Figure 5. Phrasal recognition significantly outperforms detection of participating objects and then modeling their interactions. This figure shows examples of visual phrase detections where independent objects couldn't be found using state of the art object models. For example, in image "a", the person detector failed to localize the lady in the red dress while our "person next to bicycle" detector localizes her accurately. In image "b", the person detector fails to localize the baby and our "person drinking from bottle" detector correctly finds this visual phrase.

References

[1] Y. Amit and A. Trouvé. Pop: Patchwork of parts models for object recognition. *Int. J. Comput. Vision*, 2007. 1347

[2] C. F. C. Desai, D. Ramanan. Discriminative models for multi-class object layout. In *ICCV*, 2010. 1348, 1349, 1351, 1352, 1353

[3] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001. 1347

[4] J. Coughlan, A. Yuille, C. English, and D. Snow. Efficient optimization of a deformable template using dynamic programming. *CVPR*, 1998. 1347

[5] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. *CVPR*, 2005. 1347

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 2010. 1348

[7] A. Farhadi, S. M. M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. A. Forsyth. Every picture

	bicycle	bottle	car	chair	dog	horse	person	sofa
detectors of [8]	0.434	0.429	0.329	0.213	0.316	0.438	0.295	0.204
[2] without phrases	0.431	0.425	0.191	0.225	0.297	0.475	0.204	0.167
[2] with phrases	0.449	**0.435**	0.228	0.217	0.316	0.462	0.286	0.204
Our decoding without phrases	0.437	0.434	0.330	0.216	0.329	0.440	0.297	0.218
Our decoding with phrases	**0.457**	**0.435**	**0.344**	**0.227**	**0.335**	**0.485**	**0.302**	**0.260**

Table 2. Phrasal recognition helps object detection. This table compares the performance of our decoding with that of [2] with and without visual phrases using per class AP's. Adding visual phrases helps detection of objects. This table also shows that our decoding outperforms the state of the art object detectors of [8] and state of the art multiclass recognition method of [2].

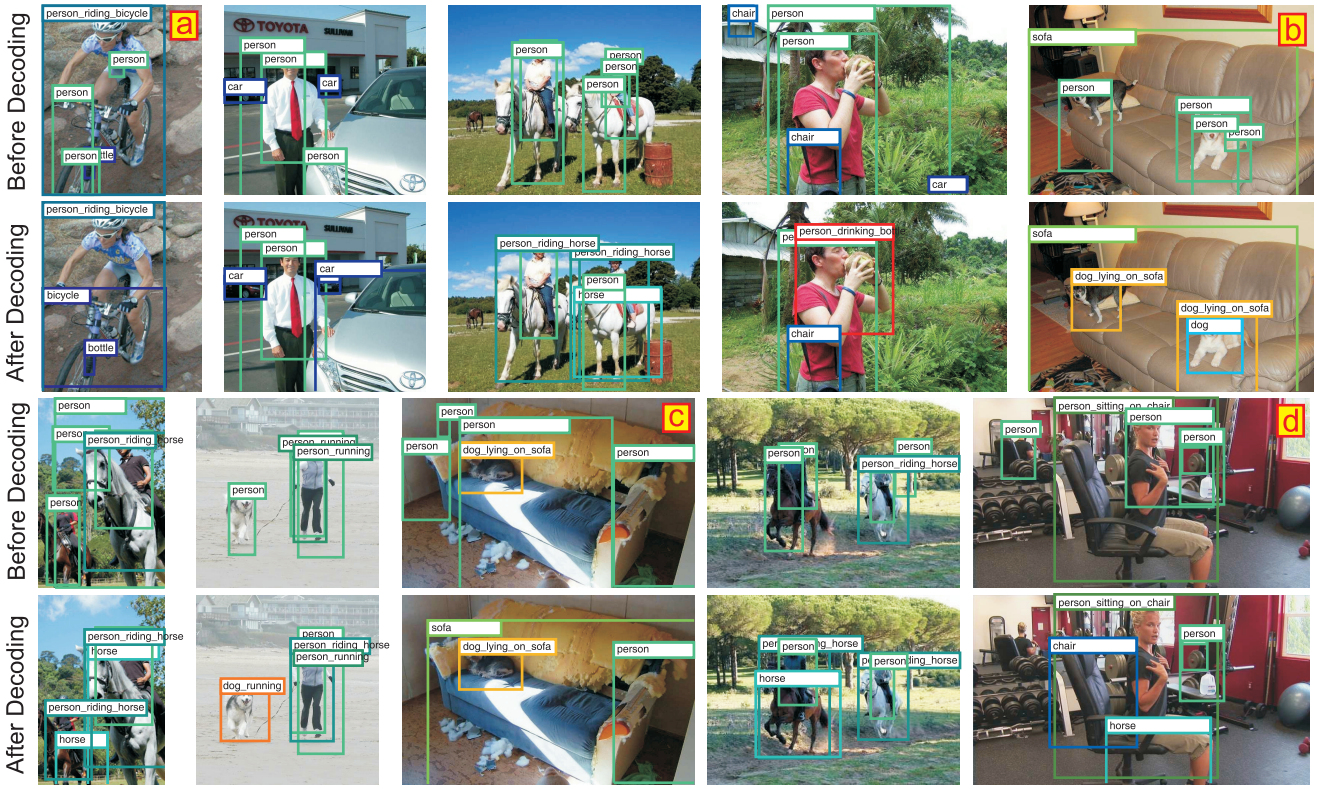

Figure 6. Rows 1 and 2 depicts our results before and after decoding, respectively. The same applies to rows 3 and 4. For example, in image "a", our decoding boosts the confidence of the bicycle classifier and suppresses the confidences of wrong person detections using a reliable "person riding bicycle" detection. In image "c", a confident "dog lying on sofa" detector improves the confidence of the sofa detection and decreases the confidences of wrong person detections. In image "d", the "person sitting on chair" detector increases the confidence of the chair detection. Our decoding shows that visual phrases help object detection and vice versa. In image "b", the confident sofa detection boosts the confidence of "dog lying on sofa" detection.

tells a story: Generating sentences from images. In *ECCV*, 2010. 1348

[8] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. http://people.cs.uchicago.edu/ pff/latent-release4/. 1349, 1350, 1351, 1353

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010. 1347, 1350, 1351, 1352

[10] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. *CVPR*, 2003. 1347

[11] A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*, 2008. 1347

[12] P. Koehn. *Statistical Machine Translation*. Cambridge University Press, 2010. 1348

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 1348

[14] N. Loeff and A. Farhadi. Scene discovery by matrix factorization. In *ECCV*, 2008. 1348

[15] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 2001. 1348

[16] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 1347