This ICCV2013 paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

A Unified Probabilistic Approach Modeling Relationships between Attributes and Objects

Xiaoyang Wang and Qiang Ji
Rensselaer Polytechnic Institute
110 Eighth Street, Troy, NY USA 12180
{wangx16,jiq}@rpi.edu

Abstract

This paper proposes a unified probabilistic model to model the relationships between attributes and objects for attribute prediction and object recognition. As a list of semantically meaningful properties of objects, attributes generally relate to each other statistically. In this paper, we propose a unified probabilistic model to automatically discover and capture both the object-dependent and object-independent attribute relationships. The model utilizes the captured relationships to benefit both attribute prediction and object recognition. Experiments on four benchmark attribute datasets demonstrate the effectiveness of the proposed unified model for improving attribute prediction as well as object recognition in both standard and zero-shot learning cases.

Figure 1. Attributes with frequency statistics with refer to different object classes. We can see the attributes "with stripe" and "white" would be co-occurring for object "zebra", but be mutually exclusive for object "polar bear". Animal examples in this figure are chosen from the Animals with Attributes (AWA) dataset [17].

1. Introduction

Attributes are a list of semantically meaningful properties shared among different objects. Recent work (e.g. [6, 17, 24]) generally utilizes attributes as an intermediate layer descriptive representation, and has applied attributes to several interesting applications like zero-shot learning [17], description of objects [6], and object recognition [34].

Standard attribute approaches generally learn a group of attribute classifiers, one for each attribute. During testing, attributes are predicted individually through the learned attribute classifiers. However, attributes are NOT independent. Different attributes generally relate to each other statistically. Moreover, due to tremendous variations in the real scene applications, it is usually difficult to individually recognize many of the attributes. And, a low quality attribute prediction can adversely affect the subsequent object recognition. This motivates us to exploit the statistical relationships between attributes and objects, and utilize these relationships to help predict certain attributes (or objects likewise) that are hard to predict alone.

We observe there exist two types of attribute relationships, i.e. the *object-dependent relationship* and *object-independent relationship*. For example, in Figure 1, the attributes "with stripe" and "white" would be co-occurring for object "zebra", but be mutually exclusive for object "polar bear". In this example, the relationship between attributes "with stripe" and "white" is object-dependent. Also, some attributes would have the same relationships among many object classes. For instance, different types of birds should all share the attributes "feature", "wing" and "leg". The relationships among these attributes are object independent. The object-dependent relationship is resulted from specific properties of an object. The attributes must either be co-occurring or be mutually exclusive to accurately describe the object. Comparatively, the object-independent relationship captures the intrinsic properties of all or many objects.

We propose a unified probabilistic model to automatically discover and capture both the object-dependent and object-independent relationships. Our unified model is essentially a Bayesian network (BN) [21]. Both the the at-

tributes and objects are represented by random variable nodes in the unified model. During model learning, training samples consisting of ground truth attributes and object labels are used to learn the model structure and parameter. In this way, both the object-dependent and object-independent attribute relationships are automatically discovered and captured. During testing, the unified model can infer both the attribute node states and the object node state given the attribute measurements predicted individually by the pre-learned attribute classifiers. Experiments on four benchmark datasets show that the unified model can effectively improve the attribute prediction accuracy by utilizing the captured relationships. Moreover, by directly inferring the object node state, the unified model can greatly benefit the object recognition task as well.

In summary, our major contributions in this proposed work can be listed in two folds: 1) we propose to build one unified model to discover and capture both the object-dependent attribute relationships and object-independent attribute relationships simultaneously in a systematic manner; 2) with the assistance of captured relationships, our unified model can directly infer both attributes and objects, and thus benefit the attribute predictions and object recognition in both standard and zero-shot learning cases.

2. Related Work

In recent years, attributes [8] have received a lot of attentions from different areas in the computer vision field. Attributes are widely used in many different applications like object describing [6], zero-shot learning [17, 11], face verification [15], object recognition [34, 10], person and clothing describing [1, 2], image search and retrieval [16, 13, 36], and action/activity recognition [19, 35, 9] etc. These applications generally accomplish two types of tasks using attributes: 1. *describing task* which utilizes attribute predictions as a detailed object description, besides the class label; 2. *classification task* which utilizes attributes to fulfill/improve either standard or zero-shot object classification. To improve the performance on either of these two attribute related tasks, approaches can be divided into either defining/discovering better attributes, or building better models for attribute prediction and object recognition using existing attributes.

Many studies [24, 25, 13, 26, 20, 4, 36, 29, 32] focus on defining and discovering better attributes (e.g. relative attributes [25] and augmented attributes [32]). On the other hand, similar to our motivation, several approaches propose methods like multi-task feature learning [10], DAP/IAP based zero-shot learning [17], BN models [30, 5], latent SVM [34] and CRF [2] to boost the performance of object recognition or attribute prediction using existing attributes. Among these approaches, the latent SVM attribute model [34] and CRF [2] also utilize attribute relationships.

Latent SVM [7] is used by Wang *et al*. [34] for object recognition with attributes, and then applied to human action recognition in [19] and active learning of attributes in [14]. In the latent SVM based attribute approaches, attributes are treated as latent variables for both model training and testing. Also, the attribute relations [34] in the latent SVM attribute approach are either manually specified, or pre-learned through a network purely consisting of attribute nodes and then interpolated into latent SVM. By contrast, in our unified model, the semantic attributes are not latent during training of the unified model, and thus our model explicitly captures the semantic meanings of the attributes which can generalize to other classes for the zero-shot learning. Moreover, our unified model automatically discovers and captures the attribute relationships in a systematic manner, considering both the object-dependent and object-independent attribute relations.

Recently, Chen *et al*. [2] apply a CRF model to improve attribute predictions for the clothing appearance describing tasks. This CRF model consists of only attribute nodes with corresponding attribute observations. It incorporates the attribute relations by specifying a fully connected network connecting all its attribute nodes. Differently, in our unified model, we build links not only among attributes, but also between attributes and object nodes. In this way, we can capture both the object-dependent and object-independent attribute relationships. Our model can then improve the tasks of both attribute classification and object recognition either in standard or zero-shot learning scenarios. Moreover, instead of specifying a fully connected graphical model, we try to discover necessary statistical relationships between the attributes nodes and object class nodes using the incorporated structure and parameter learning.

The BN based attribute approaches by Scheirer *et al*. in [30] and Farhadi *et al*. in [5] also use BN to combine attributes. However, these two approaches do NOT incorporate attribute relationships in their models. Comparatively, discovering, capturing, and exploiting attribute relation is the essence of our paper. Thus, the BN approaches in [30] and [5] are very different from our approach.

Another probabilistic attribute approach is the direct attribute prediction (DAP) and indirect attribute prediction (IAP) models proposed by Lampert *et al*. [17] for zero-shot based recognition of new objects. Even though DAP and IAP are probabilistic approaches, they are different from our unified model approach. The DAP and IAP models assume the attribute vector can be induced deterministically given the class label, but our model assumes that given the class label, the related attributes still obey certain probabilistic distributions. Also, our unified model discovers and captures the object-dependent and object-independent attribute relationships, but DAP and IAP models do not capture such relationships.

3. Approach

In Section 3, we describe in details about the unified model we propose to capture the relationships between attributes and objects for attribute prediction and object recognition. In our approach, we denote Y as object label, $A_1, A_2, \ldots, A_M$ as the M ground truth attributes, and X as the raw feature shared by attribute classifiers.

3.1. Modeling Relationships

The *object-dependent* attribute relationships are the relationships resulted from specific properties of an object. Comparatively, the *object-independent* attribute relationships capture intrinsic properties among all or many objects. We believe attribute relationships should consist of both types. To automatically differentiate and capture these relationships, we construct a Bayesian network (BN) based unified model consisting of both object label node Y and attribute nodes $A_1, A_2, \ldots, A_M$. A BN is a directed acyclic graph (DAG) which represents a joint probability distribution among a set of variables [21], where the nodes denote random variables and the links denote the conditional dependencies among variables. Figure 2 gives an example of a BN capturing the relationships among attributes $A_1, A_2, \ldots, A_M$ and object Y.

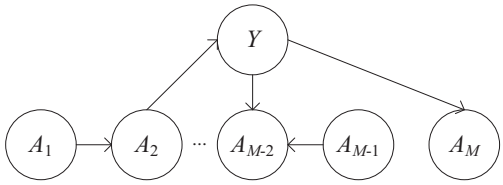

Figure 2. A BN capturing the relationships among attributes $A_1, A_2, \ldots, A_M$ and object category Y.

In our unified model as shown in Figure 2, attribute nodes connecting to each other without connecting to object node capture *object-independent* relationships, while attribute nodes connecting to each other via object node capture *object-dependent* relationships. Compared to undirected graphical models (CRFs, MRFs etc.), BN can easily learn its optimal structure directly from data. This advantage of BN enables us to discover the attribute relationships systematically in the model learning phase.

3.2. Learning Model Structure

In the BN model shown in Figure 2, the node Y represents the object label, and the M nodes $A_1, A_2, \ldots, A_M$ denote the M attributes. The directed links between nodes Y and $A_1, A_2, \ldots, A_M$ form a DAG that captures the structure $\mathcal{G}$ of the BN. To learn such structure $\mathcal{G}$ from training data D (the ground truth attributes and the object labels of training samples), we use Bayesian information criterion (BIC) [31] as the score to evaluate the fitness of a possible structure $\mathcal{G}_s$:

$$Score(\mathcal{G}_s : D) = \log P(D|\hat{\theta}_{\mathcal{G}_s}, \mathcal{G}_s) - \frac{d(\hat{\theta}_{\mathcal{G}_s})}{2} \log N \quad (1)$$

where $\hat{\theta}_{\mathcal{G}_s}$ is the estimation of model parameter with structure $\mathcal{G}_s$, $d(\hat{\theta}_{\mathcal{G}_s})$ is the number of free parameters in $\hat{\theta}_{\mathcal{G}_s}$, and N is the number of samples in training data D.

In Equation 1, the first term on right represents the joint log-likelihood of data D with the possible model structure $\mathcal{G}_s$ and the corresponding parameter $\hat{\theta}_{\mathcal{G}_s}$. It evaluates how well the network $\mathcal{G}_s$ fits the data. The second term is a penalty term. It is proportional to the number of free parameters reflecting the complexity of the network.

To find the global optimal structure $\mathcal{G}$ that maximizes the BIC score $Score(\mathcal{G}_s : D)$ from the set of possible structures $\mathcal{G}_s$, we employ the structure learning method proposed in [3], where branch-and-bound has been applied to perform the exact learning of BN structure. To avoid a complex BN with too many parameters, we limit the number of parental nodes for each attribute, based on the assumption that each attribute is only closely related to a few (N) attributes. We empirically set $N = 3$ to achieve an optimal tradeoff between structure complexity and recognition performance.

3.3. Learning Model Parameters

Parameters for our unified model involve the conditional probability table (CPT) for each node given its parents. The usual way of learning such parameters is to use the maximum likelihood estimation (MLE). However, for several cases in the unified model, a certain parent-child state combination would seldom appear (e.g. the object label state representing "sheep" with the attribute "black" to be true), and the MLE learning of such parameter would be affected by the limited training samples for the certain state parent-child combination. Hence, we use the maximum a posteriori (MAP) estimation to learn the model parameters instead.

Given a training set D containing object labels, ground truth attributes and their corresponding measurements, our goal is to estimate the parameter θ for a BN with structure $\mathcal{G}$ by MAP approach as shown in Equation 2.

$$\theta^* = \arg\max_{\theta} P(\theta|D, \mathcal{G}) = \arg\max_{\theta} P(D|\theta, \mathcal{G}) P(\theta) \quad (2)$$

More specifically, for the discrete node network, we denote one of the discrete nodes in the unified model to be X_i. For the parameter learning of this discrete node X_i, the posterior distribution of parameter θ_{ij} for X_i given its parent(s) $pa(X_i)$ in state j can be represented by Dirichlet distribution:

$$P(\theta_{ij}|D, \mathcal{G}) = Dir(\alpha_{ij1} + N_{ij1}, \ldots, \alpha_{ijr_i} + N_{ijr_i}) \quad (3)$$

where r_i is the number of states for variable X_i. We use $X_i = k$ to denote that node X_i is in state k, and use

$pa(X_i) = j$ to denote the parent(s) of node X_i to be in state j. And, N_{ijk} reflects the number of cases in the training set D for which $X_i = k$ and $pa(X_i) = j$, and α_{ijk} is the hyper-parameter that reflects the prior beliefs about how often the case $X_i = k$ and $pa(X_i) = j$ would appear.

With the MAP approach in Equation 2, the analytical solution for parameter θ_{ijk} of node X_i which stands for the probability $P(X_i = k|pa(X_i) = j)$ would be:

$$\theta_{ijk} = \frac{\alpha_{ijk} + N_{ijk} - 1}{\alpha_{ij} + N_{ij} - r_i} \quad (4)$$

where $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$, and $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$.

3.4. Inferring Attributes and Object Classes

Based on the learned unified model with structure $\mathcal{G}$ and parameter θ, both the attribute prediction and object recognition can be performed through the model inference. The independent attribute classifiers pre-trained on the training images are used first to obtain the attribute measurements $OA_1, OA_2, \ldots, OA_M$ from testing image X. And then, these attribute measurements are used by the unified model for inferring the attributes and object classes.

To incorporate these attribute measurements into the unified model for testing, we further associate the attribute nodes $A_1, A_2, \ldots, A_M$ each with a measurement node. Figure 3 shows an example of the unified model incorporated with attribute measurements. In Figure 3, the object label node Y and attribute nodes $A_1, A_2, \ldots, A_M$ are indicated with white circles, and the attribute measurement nodes $OA_1, OA_2, \ldots, OA_M$ are indicated with shaded circles. The links between the attribute nodes and the attribute measurement nodes model the measurement uncertainty of the independent attribute classifiers. From the BN model point of view, the measurement nodes are regarded as observed nodes, which provide evidence in the inference procedure, and the object label and attribute nodes are ground truth nodes, whose states need to be inferred from the BN model given the evidence.

The factorized form of the joint probability for the unified model incorporated with attribute measurements is:

$$P(Y, A_1, \ldots, A_M, OY, OA_1, \ldots, OA_M) = P(Y|pa(Y)) \prod_{m=1}^{M} P(A_m|pa(A_m)) \prod_{m=1}^{M} P(OA_m|A_m) \quad (5)$$

where $pa(Y)$ stands for the parent node(s) of object label node Y, and $pa(A_m)$ stands for the parent node(s) of attribute node A_m. Terms $P(Y|pa(Y))$ and $P(A_m|pa(A_m))$ represent the conditional dependencies among object label node and attribute nodes, which capture the object-dependent and object-independent attribute relationships, and the term $P(OA_m|A_m)$ represents the attribute measurement uncertainty terms. In practice, we use discrete

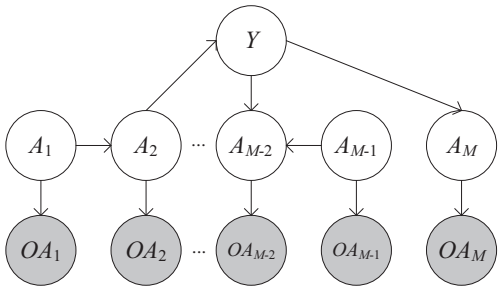

Figure 3. A unified model incorporating attribute measurements for attribute predictions and object recognition. The object label node Y and attribute nodes $A_1, A_2, \ldots, A_M$ are indicated with white circles. The attribute measurement nodes $OA_1, OA_2, \ldots, OA_M$ are indicated with shaded circles.

measurements for OA_m, and $P(OA_m|A_m)$ is therefore in multinomial distribution.

With the unified model incorporated with attribute measurements shown in Figure 3, we can infer the probabilities of different attributes and object classes given the attribute measurements obtained from independent attribute classifiers. For object recognition, given a testing image X with its attribute measurements $OA_1, OA_2, \ldots, OA_M$, we infer the marginal probability of Y given the attribute measurements. The classification output c should be:

$$c = \arg\max_k P(Y = k|OA_1, \ldots, OA_M; \theta, \mathcal{G}') \quad (6)$$

Also, for attribute prediction, we infer the marginal probability of A_m with $m = 1, 2, \ldots, M$ given the measurements of all attributes. The attribute prediction a_m is:

$$a_m = \arg\max_r P(A_m = r|OA_1, \ldots, OA_M; \theta, \mathcal{G}') \quad (7)$$

Both inference problems can be solved efficiently by the junction tree inference method [18]. In addition, the attribute states and the object state can also be inferred jointly with the most probable explanation (MPE) [21] of the evidences. However, in our experiments, this inference performs not as well as the inferences in Equation 6 and 7.

3.5. Unified Model for New Objects

Attributes are an ideal type of semantic knowledge for zero-shot based recognition of new object classes [23, 17]. In the zero-shot setting, no raw images of new objects are available during model learning. However, with our unified model, we still want to discover and capture the statistical relationships between attributes and new objects even without raw images during training. Inspired by the real-valued association strength between attributes and classes given in [17] that is averaged from responses of 10 test persons, we learn these relationships directly from the *semantic knowledge base* [23] consisting of only new class labels and

their ground truth semantic attributes. In this way, our unified model is built purely with semantic knowledge. On the other hand, the independent attribute classifiers are learned with raw images from observed classes.

During testing, the independent attribute classifiers are applied to raw images of new classes to obtain attribute measurements, which are further applied to unified model for inference. We still use the inferences discussed in Equation 7 and 6 for attribute predictions in new class examples and the classification of new classes respectively.

4. Experiments

We demonstrate the effectiveness of our method with the following four different vision datasets.

The **a-Pascal** dataset [6] contains 6340 training samples and 6355 testing samples collected from Pascal VOC 2008 challenge. Each sample belongs to one of the twenty object classes. The **a-Yahoo** dataset [6] contains 2644 samples belonging to twelve classes that are completely different from the classes in a-Pascal dataset. A list of 64 attributes are provided for each sample of both a-Pascal and a-Yahoo datasets. These datasets also provide a 9751 dimension base feature for each of the training and testing samples. A support vector machine (SVM) based attribute classifier trained on the training set of a-Pascal is provided in [6] to predict the 64 attributes with the given features.

The **SUN Attribute** dataset collected by Patterson *et al*. [28, 27] contains a pool of 14,340 scene images belonging to 717 scene classes. A set of 102 manually labeled attributes is available for each of the images. Instead of using random splits of the training and testing images in [28], Patterson *et al*. recently update the SUN attribute dataset (v2.1) in [27] by specifying the fixed training and testing sets with 12906 and 1434 images respectively. The pre-calculated image features and attribute classifiers are provided in [27].

The **Animals with Attributes** (AWA) dataset [17] consists of 30475 animal images belong to 50 classes. Among these 50 classes of animals, 10 classes are selected for testing. The 6180 images belonging to these 10 classes act as test data, and the 24295 images of remaining classes are used for training. The real-valued association strength [22, 12] between 85 attributes and 50 animal classes are provided together with the binary definitions of attributes. This dataset also provides pre-calculated feature vector and the baseline attribute classifiers.

4.1. Discovering and Capturing Relationships

We first show the structure learning result for discovering and capturing the relationships between attributes and objects. To demonstrate a distinct structure for analysis, we experiment on the two objects "car" and "bicycle" of the a-Pascal dataset[1]. 14 attributes are involved for these two objects: headlight, taillight, handlebar, exhaust, text, wheel, metal, plastic, window, door, side mirror, pedal, 3D Boxy and shinny. The learned global optimal structure is presented in Figure 4.

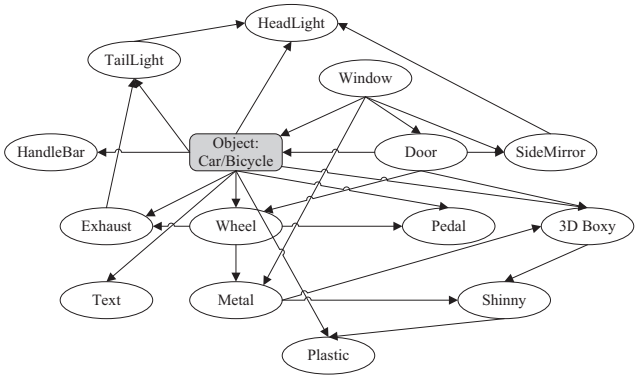

Figure 4. The learned structure capturing relationships between attributes and objects. This structure is learned with ground truth attributes of two object classes "car" and "bicycle" on a-Pascal. The shaded rectangle node represents the object node. The elliptical nodes represent the attribute nodes.

The learned structure shown in Figure 4 captures many of the relationships. For example, the relationship between attributes "exhaust" and "wheel" is the object-dependent relationship (both car and bicycle should have wheels, but bicycle should not have exhaust and car should have), and the learned structure indicates the distribution of attribute node "exhaust" would depend on the states of both the object node and the attribute node "wheel".

4.2. Performance for Attribute Prediction

Attribute prediction can provide more detailed semantic descriptions about the target. Here, we first test our unified models for attribute prediction both with observed classes on a-Pascal dataset [6], and with new classes on a-Yahoo dataset [6] in Section 4.2.1. Then, we compare with existing results on SUN attribute dataset [28, 27] in Section 4.2.2.

4.2.1 Attribute Prediction on a-Pascal and a-Yahoo

In this experiment, we use the proposed unified model for attribute prediction of observed class images on a-Pascal [6] and new class images on a-Yahoo respectively with the inference discussed in Equation 7. We use the provided attribute classifiers [6] trained on the a-Pascal training set as "Baseline" for attribute prediction, and these baseline attribute classifiers also provide attribute measurements for our unified model to predict attributes. The Geometric Mean (G-mean) [2] is used for prediction accuracy evaluation in the purpose of further comparing with the CRF

[1] Structure learned with all objects is in supplementary material.

in [2]. Figure 5 and 6 give the per-attribute G-Mean comparison between the Baseline and the proposed unified model for attribute prediction of observed classes on a-Pascal and new classes on a-Yahoo respectively.

From Figure 5, we can see the proposed unified model can improve the attribute prediction over the attribute measurements for most of the attributes on a-Pascal dataset. For certain attributes like "sail" and "saddle", the improvement is very significant. The G-Mean for attribute "sail" improves from 55.18% to 86.31%. Also, in Figure 6, the improvements for attribute prediction in images of new classes on a-Yahoo dataset are even greater than that in a-Pascal. We believe the captured object-dependent and object-independent attribute relationships in a-Yahoo classes significantly benefit the attribute prediction for these new class images, even though the baseline attribute classifiers perform worse on a-Yahoo due to generalization issues.

We further compare the baseline independent attribute classifier and the proposed unified model on the average G-mean over all 64 attributes in Table 1. To verify the effectiveness of the distinction between object-(in)dependent attribute relations, we compare with a BN model consisting of attribute nodes but omitting the object node (BN-Att model). This BN-Att model is learned in the same procedure as the proposed model. Also, the CRF in [2] using a fully connected network of only attribute nodes to relate attributes is also compared here. From this comparison, we can see that incorporating the traditional attribute relationships alone by the BN-Att or CRF model can already improve the baseline performance. Moreover, with our proposed model that differentiates the object-dependent and object-independent attribute relationships, the overall attribute prediction accuracy can be significantly improved. In addition, for our proposed model, the overall improvement is again more significant for the a-Yahoo database, demonstrating the generalization ability of our approach.

Table 1. Overall attribute prediction accuracy w.r.t different models on a-Pascal test set and a-Yahoo.

	Baseline	BN-Att	CRF [2]	Proposed Model
a-Pascal	70.41%	72.42%	74.01%	79.03%
a-Yahoo	65.40%	66.50%	66.81%	78.12%

4.2.2 Attribute Prediction on SUN Attribute Dataset

To study model performance for complex cases that involve large attribute set, we also perform attribute prediction on the SUN Attribute dataset. This dataset contains scene images belonging to 717 scene classes. And, 102 attributes are defined in total. Most recently, Patterson *et al*. release the latest attribute prediction results in [27] with the new training and testing splits. In [27], the average precision (AP) number is used for evaluating attribute prediction accuracy. And, the mean average precision (Mean AP) over all 102 attributes is 50.22% in [27]. For comparison, we also use Mean AP for evaluation. For attribute prediction, our model improves the overall result from 50.22% in [27] to 51.12% on the Mean AP evaluation over all 102 attributes. It takes 48.8 min for our proposed model to recognize all 102 scene attributes for 1434 samples on an Intel i7 2.93GHz computer. This translates to 2.04 sec for recognizing each sample, and 0.02 sec for recognizing each attribute.

4.3. Performance for Object Recognition

Attributes can generally be utilized to help the task of object recognition, especially in zero-shot learning cases as in [17] where image examples of new classes are not available during all phases of model learning. We utilize the proposed unified model for the object recognition of observed classes on a-Pascal dataset in Section 4.3.1, and of new classes on a-Yahoo and AWA datasets in Section 4.3.2.

4.3.1 Object Recognition for Observed Classes

We test the proposed unified model for recognizing the observed object classes using the a-Pascal dataset [6]. Both the attribute classifiers and the unified model are learned on the a-Pascal training set. The overall performance on the a-Pascal testing set is given in Table 2. Both the mean per-class and the overall accuracies are used for evaluation.

Table 2. Object recognition with attributes on a-Pascal dataset.

	SVM [6]	Farhadi *et al*. [6]	Wang *et al*. [34]	Proposed Model
Mean Per-class	35.5%	37.7%	50.84%	44.82%
Overall	58.5%	59.4%	59.15%	63.02%

In Table 2, the SVM approach given in [6] uses the base feature directly for training and testing. With the help of attributes, the approach proposed by Farhadi *et al*. [6] can already outperform the SVM built on the base feature. Our proposed model can outperform the approach in Farhadi *et al*. [6] on both the mean per-class and overall evaluations. Compared to the approach by Wang *et al*. [34], our model performs better in overall accuracy, but not as well in the mean per-class accuracy. This is expected, since [34] use a loss function specifically designed for skewed data, and the a-Pascal dataset is skewed (2571 of the 6355 test samples are "person"). The result of [34] with standard "0/1" loss function is 46.25% for mean per-class and 62.16% for overall, and our result is close to this performance.

Here, we further compare with the per-object BN network where a BN model consisting of only attribute nodes is learned for each object class. For testing, object label is predicted by picking the model with the maximum likelihood.

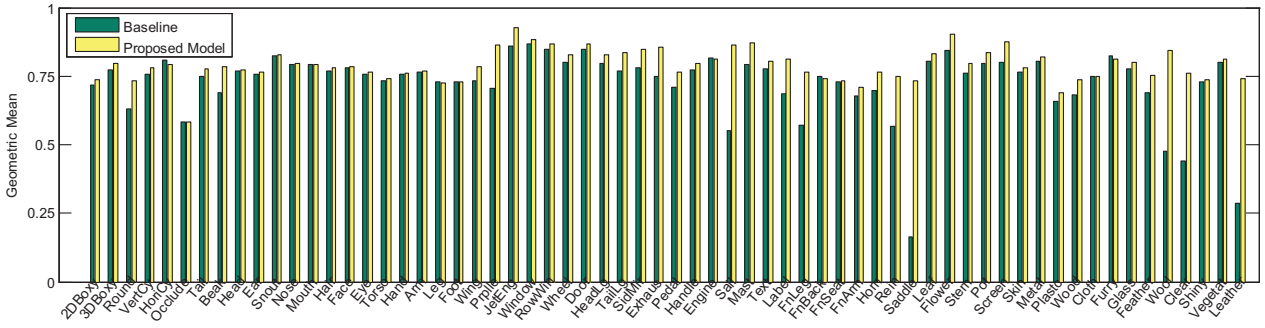

Figure 5. Accuracies of attribute predictions on a-Pascal test set. The proposed unified model can improve the attribute prediction over the measurement for most of the attributes, with more significant improvements for certain attributes like "sail" and "saddle".

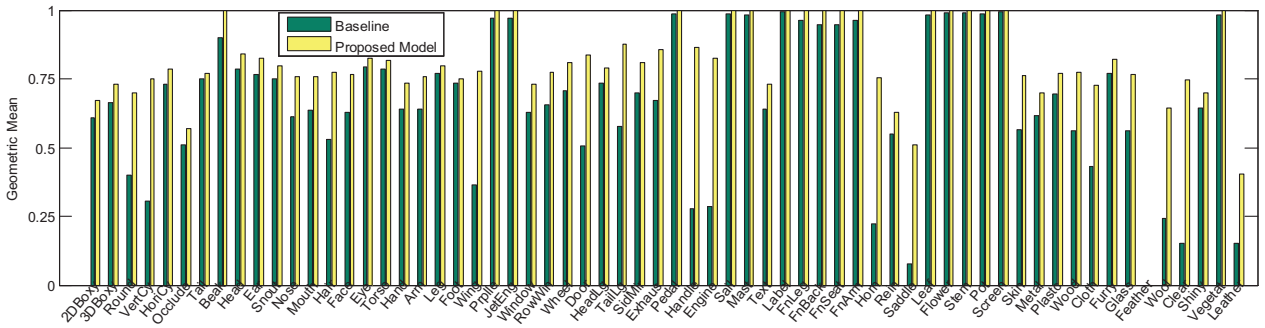

Figure 6. Accuracies of attribute predictions on a-Yahoo dataset for images of new classes. By using our proposed model, the predictions for most of the attributes are improved. For attributes like "wing", "handle" and "engine", the improvements are very significant.

This per-object network achieved 42.09% and 60.24% for average and overall accuracy, lower than proposed model.

4.3.2 Object Recognition with Zero-shot Learning

We first test the proposed unified model for recognizing new classes on a-Yahoo dataset [6] in zero-shot learning scenario. The attribute classifiers are trained on a-Pascal training set and provide attribute measurements on a-Yahoo images as input for unified model. We choose SVM built on the same *semantic knowledge base* to serve as the baseline object recognition model. During testing, SVM uses interpreted attributes predicted both from the attribute classifier (i.e. original attribute predictions), and from the unified model (i.e. improved attribute predictions by our unified model in Section 4.2.1) for a-Yahoo object recognition.

Table 3. Results for recognizing new object classes with zero-shot learning on a-Yahoo dataset.

	SVM[a]	SVM[b]	Farhadi *et al.* [6]	Proposed Model
Mean Per-class	17.39%	33.74%	N/A	41.31%
Overall	16.38%	39.49%	32.5%	**45.05%**

SVM[a]: testing on original attribute predictions; SVM[b]: testing on improved attribute predictions by our unified model.

Table 3 gives the overall evaluations for SVM testing on original attribute predictions as well as testing on attribute predictions by our unified model. We can see the improved attribute predictions can also significantly benefit the object recognition on new classes. However, the best performance is reached by directly inferring the object state through the proposed unified model as discussed in Equation 6. We also compare our proposed model result with the result given by Farhadi *et al.* [6] in the approach "learning new categories from textual description". This approach in [6] is also a zero-shot learning approach that recognizes new classes whose image examples are omitted from the training set. The mean per-class accuracy for this approach is not available in [6], but our proposed model can improve the overall recognition accuracy by over 12%.

Also, Lampert *et al.* [17] propose two attribute based zero-shot learning models DAP and IAP. We further compare with these two models on the AWA dataset [17]. We test on the same 10 animal classes as defined in [17], and use the provided baseline attribute classifiers trained on the images of the rest 40 animals to obtain attribute measurement input. The comparisons are shown in Table 4, where the mean-per class recognition rates of DAP and IAP models are from [17] directly, and the overall recognition rates of DAP and IAP models are calculated from the given confusion matrices in the dataset. In this comparison, our unified model can outperform both IAP and DAP models for attribute based zero-shot learning.

For experiments in Section 4.3.2, several other approaches (e.g. [34, 10]) that also use the a-Yahoo or AWA

Table 4. Comparison with DAP and IAP models for recognizing new object classes with zero-shot learning on AWA dataset.

	Lampert *et al.* - IAP [17]	Lampert *et al.* - DAP [17]	Proposed Model
Mean Per-class	27.8%	40.5%	**43.36%**
Overall	29.69%	39.74%	**42.78%**

dataset are not comparable with our approach since those approaches use the image examples of the new classes directly or indirectly for learning the object classifiers. More specifically, the approach in [34] randomly split the a-Yahoo images into training/testing sets and use the images of a-Yahoo classes directly during training; in the 10-class AWA subset experiment of [10], the object classifier is trained directly on the image input of the selected 10 classes, and regularized by the attribute classifiers trained on different sets of images; the "learning to identify new objects" approach in [6] utilizes the predicted attributes on a-Yahoo as training data to train object classifiers, and thus in turn uses the a-Yahoo images indirectly for classifier training in the classifier cascade. Comparatively, these approaches can generally achieve around 70% accuracy on the corresponding dataset by using the image examples of new classes directly or indirectly.

5. Conclusion

In this paper, we propose a unified probabilistic model to capture the relationships between attributes and objects for attribute prediction and object recognition. As a list of semantically meaningful properties of objects, attributes generally relate to each other statistically. The paper proposes a unified probabilistic model to automatically discover and capture both the object-dependent and object-independent attribute relationships. During testing, the unified model utilizes these captured relationships to benefit both attribute prediction and object recognition with probabilistic inference given the attribute measurements predicted individually by the pre-learned attribute classifiers. We experiment on four benchmark attribute datasets including a-Pascal, a-Yahoo, SUN Attribute and AWA for attribute prediction and object recognition tasks. The experiment results with the proposed unified model show significant improvements for attribute prediction as well as object recognition, especially in cases of new objects.

Acknowledgments

This work is funded in part by US Defense Advanced Research Projects Agency under grants HR0011-08-C-0135-S8 and HR0011-10-C-0112.

References

[1] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, 2011. 2

[2] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*, 2012. 2, 5, 6

[3] C. P. de Campos, Z. Zeng, and Q. Ji. Structure learning of bayesian networks using constraints. In *ICML*, 2009. 3

[4] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*, 2012. 2

[5] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *CVPR*, 2010. 2

[6] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1, 2, 5, 6, 7, 8

[7] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 2

[8] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*. 2007. 2

[9] Y. Fu, T. Hospedales, T. Xiang, and S. Gong. Attribute learning for understanding unstructured social activity. In *ECCV*, 2012. 2

[10] S. J. Hwang, F. Sha, and K. Grauman. Sharing features between objects and their attributes. In *CVPR*, 2011. 2, 7, 8

[11] P. Kankuekul, A. Kawewong, S. Tangruamsub, and O. Hasegawa. Online incremental attribute-based zero-shot learning. In *CVPR*, 2012. 2

[12] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, 2006. 5

[13] A. Kovashka, D. Parikh, and K. Grauman. Whittle search: Image search with relative attribute feedback. In *CVPR*, 2012. 2

[14] A. Kovashka, S. Vijayanarasimhan, and K. Grauman. Actively selecting annotations among objects and attributes. In *ICCV*, 2011. 2

[15] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 2

[16] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *PAMI, IEEE Transactions on*, 33(10):1962 –1977, 2011. 2

[17] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1, 2, 4, 5, 6, 7, 8

[18] S. Lauritzen and D. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *J. Royal Statistics Society B*, pages 157–194, 1988. 4

[19] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011. 2

[20] D. Mahajan, S. Sellamanickam, and V. Nair. A joint learning framework for attribute models and object descriptions. In *ICCV*, 2011. 2

[21] K. P. Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, 2012. 1, 3, 4

[22] D. N. Osherson, J. Stern, O. Wilkie, M. Stob, and E. E. Smith. Default probability. *Cognitive Science*, 15(2):251–269, 1991. 5

[23] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*. 2009. 4

[24] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, 2011. 1, 2

[25] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011. 2

[26] A. Parkash and D. Parikh. Attributes for classifier feedback. In *ECCV*, 2012. 2

[27] G. Patterson, X. Chen, and J. Hays. Sun attribute database. `http://cs.brown.edu/~gen/sunattributes.html`. 5, 6

[28] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2012. 5

[29] M. Rastegari, A. Farhadi, and D. Forsyth. Attribute discovery via predictable discriminative binary codes. In *ECCV*. 2012. 2

[30] W. J. Scheirer, N. Kumar, K. Ricanek, P. N. Belhumeur, and T. E. Boult. Fusing with context: a bayesian approach to combining descriptive attributes. In *IJCB*, 2011. 2

[31] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978. 3

[32] V. Sharmanska, N. Quadrianto, and C. Lampert. Augmented attribute representation. In *ECCV*, 2012. 2

[33] X. Wang and Q. Ji. A novel probabilistic approach utilizing clip attributes as hidden knowledge for event recognition. In *ICPR*, 2012.

[34] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *ECCV*. 2010. 1, 2, 6, 7, 8

[35] B. Yao, X. Jiang, A. Khosla, A. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011. 2

[36] F. Yu, R. Ji, M.-H. Tsai, G. Ye, and S.-F. Chang. Weak attributes for large-scale image retrieval. In *CVPR*, 2012. 2