Object Recognition with Hidden Attributes

Xiaoyang Wang and Qiang Ji
Rensselaer Polytechnic Institute, Troy, NY, USA
xiaoyang.wangs@gmail.com, jiq@rpi.edu

Abstract

Attribute based object recognition performs object recognition using the semantic properties of the object. Unlike the existing approaches that treat attributes as a middle level representation and require to estimate the attributes during testing, we propose to incorporate the hidden attributes, which are the attributes used only during training to improve model learning and are not needed during testing. To achieve this goal, we develop two different approaches to incorporate hidden attributes. The first approach utilizes hidden attributes as additional information to improve the object classification model. The second approach further exploits the semantic relationships between the objects and the hidden attributes. Experiments on benchmark data sets demonstrate that both approaches can effectively improve the learning of the object classifiers over the baseline models that do not use attributes, and their combination reaches the best performance. Experiments also show that the proposed approaches outperform both state of the art methods that use attributes as middle level representation and the approaches that learn the classifiers with hidden information.

1 Introduction

Object recognition in computer vision generally refers to the recognition of object images into different categories such as "bird", "aeroplane", "bicycle", etc. In recent years, computer vision researchers explore to assign a list of attributes [Farhadi *et al.*, 2009] to the object images. These attributes [Ferrari and Zisserman, 2007] are manually specified and semantically meaningful descriptions about the object shape (e.g. "is cylindrical"), parts (e.g. "has head", "has leg"), materials (e.g. "made of wood"), color (e.g. "is red"), etc. The approaches [Farhadi *et al.*, 2009; Lampert *et al.*, 2009; Wang and Mori, 2010; Parikh and Grauman, 2011a; Kovashka *et al.*, 2011]) that utilize the assigned attributes to benefit object recognition can be called attribute-based object recognition.

Most existing attribute-based object recognition approaches (e.g. [Farhadi *et al.*, 2009; Lampert *et al.*, 2009; Wang and Mori, 2010; Parikh and Grauman, 2011a; 2011b; Kovashka *et al.*, 2011]) utilize attributes as an intermediate layer in the classifiers cascade. In the testing phase of these approaches, attributes are first predicted by the pre-trained attribute classifiers. And, these predicted attributes are further utilized by the attribute based object classifiers as mid-level input for object recognition. Typical applications of these approaches include zero-shot transfer learning [Lampert *et al.*, 2009], description of unfamiliar objects [Farhadi *et al.*, 2009], and improving the object classification [Wang and Mori, 2010], event recognition [Wang and Ji, 2012], and phone recognition [Zhao *et al.*, 2015].

Figure 1: An example of attributes for two animals, where the bird has "beak", "wing" and is covered with "feather", and the cow has "ear", "snout" and is "furry".

However, due to tremendous variations in vision applications, attribute recognition itself is challenging. Moreover, poor quality attribute measurements in the middle level would adversely affect the subsequent object classification. This dilemma motivates us to avoid utilizing attributes as middle level representation, and to explore incorporating attributes in a different way, where we have access to ground truth attributes during training, but do not utilize the predicted attributes explicitly or implicitly for final stage recognition during testing. In this paper, we call these attributes that are available only during training as hidden attributes. We hope the hidden attributes utilized in our approach can still improve the object recognition.

This hidden attribute setting lies in the "learning with hidden information" (LHI) paradigm [Vapnik and Vashist, 2009]. In this paradigm, the hidden information a are utilized only during training. They help learn a better classifier (e.g. linear classifier $y = \text{sign}(w^\top x)$) from feature x to label y that can outperform the traditional classifier (e.g. $y = \text{sign}(w_0^\top x)$) learned without hidden information. Hence, in this paradigm, the hidden information serves only for the purpose of obtaining a better parameter vector w that is in the

same dimension as the original parameter vector w_0. Compared to the traditional mid-level based attribute approaches, this paradigm can avoid the propagation of erroneous attribute predictions to the subsequent object classification.

In this paper, we propose two novel formulations to incorporate hidden attributes during model learning. Our first approach (xLR+) utilizes hidden attributes as additional information to improve the target model that predicts object label y from image feature x. In the second approach (LR-Rel+), we further incorporate the semantic relationships between the objects and the hidden attributes. Finally,we further combine both formulations as regularization terms into one unified learning objective (xLR-Rel+) that receives the best performance.

In summary, the major contributions of this work include: 1) we propose to incorporate hidden attributes for object classification; 2) we propose the formulations including xLR+, LR-Rel+ and xLR-Rel+ that use hidden attributes as extra information and exploit their relationships with objects.

2 Related Work

Utilizing attributes to enhance the object recognition performance has drawn great attention in recent years. Work in this area can be divided into *sequential attribute and object recognition*, and *joint attribute and object recognition*. The *sequential* approaches [Farhadi *et al.*, 2009; Lampert *et al.*, 2009; Parikh and Grauman, 2011a; 2011b; Akata *et al.*, 2013; Wang and Ji, 2014] utilize attributes as an intermediate representation between low-level image features and high-level categories. [Farhadi *et al.*, 2009] use linear classifiers like SVM to predict attributes from shared image features, and then use the predicted attributes for object categorization. However, the sequential approaches still require to train attribute classifiers with training data, and then predict the attribute labels or infer attribute scores during testing. The performance of these methods is therefore subject to the performance of the attribute classifiers. To address this problem, the *joint* approach performs attribute and object recognition simultaneously in order to exploit their interdependencies. Wang and Ji [Wang and Ji, 2013] utilize a Bayesian network (BN) with learned structure to improve both attribute prediction and object recognition with captured attribute relationships. Also, the multi-task learning approach in [Hwang *et al.*, 2011] simultaneously learns multiple classifiers for object recognition and attribute prediction tasks based on the shared feature assumption. The joint approaches need learn attribute and object classifiers simultaneously and they are therefore computationally complex. In contrast, our approach focuses only on the object recognition task.

Recently, [Wang *et al.*, 2014] propose to incorporate hidden information for learning logistic regression classifier LR+. While its formulation looks similar to the first formulation in this paper, our work significantly differs from the work in [Wang *et al.*, 2014]. The work in [Wang *et al.*, 2014] studies learning LR+ using hidden information as extra information. Comparatively, our work focuses on object recognition with hidden attributes. We incorporate hidden attributes as extra information in xLR+, exploit the object-attribute relationships in LR-Rel+, and further propose a combined formulation xLR-Rel+. Both LR-Rel+ and xLR-Rel+ are completely different from the LR+ in [Wang *et al.*, 2014]. Our xLR+, LR-Rel+, and xLR-Rel+ approaches all outperform the LR+ approach by [Wang *et al.*, 2014] in the experiments.

3 Object Recognition With Hidden Attributes

We first define both the traditional supervised object recognition and the the proposed object recognition with hidden attributes.

Traditional supervised object recognition can generally be formulated as: given a set of N labeled training samples represented by image feature vector set $X = \{x_1, \cdots, x_N\} \subset \mathcal{X} \subset \mathbb{R}^d$, and the object label set $Y = \{y_1, \cdots, y_N\} \in \mathcal{Y}$ with $\mathcal{Y} = \{-1, 1\}$ for binary cases, learn a mapping function $f : \mathcal{X} \longmapsto \mathbb{R}$ with parameter w from the function space $\mathcal{F}$ of all possible functions (e.g. all linear functions $f(x) = w^\top x$) to predict the object label y from the input image feature x as accurate as possible. Generally, the object classifier parameter w can be learned by minimizing the objective function shown in Equation 1, where $l(y_i, x_i; w)$ is the loss function and $||w||_2^2$ is a regularization term to avoid overfitting.

$$\min_w \sum_{i=1}^{N} l(y_i, x_i; w) + \frac{\gamma}{2}\|w\|_2^2 \tag{1}$$

Object recognition with hidden attributes differs from the traditional supervised object recognition problem in that additional hidden information vectors (i.e., the ground truth attributes in this paper) $A = \{a_1, \cdots, a_N\} \subset \mathcal{A}$ are also provided for each training sample, where each M dimensional vector a_i corresponds to the training sample pair (x_i, y_i). *Object recognition with hidden attributes* can hence be stated as: given N labeled training triplets $\{(x_i, a_i, y_i)_{i=1}^N\}$, learn a mapping function $f' : \mathcal{X} \longmapsto \mathbb{R}$, with parameters w' from the same function space $\mathcal{F}$ of all possible functions (e.g. all linear functions $f(x) = w^\top x$) to predict object label y from input image feature x as accurate as possible.

Following this definition, the new mapping function $f' : \mathcal{X} \longmapsto \mathbb{R}$ does not depend on hidden attribute space $\mathcal{A}$, but hidden attributes will influence parameter w' in training. We expect w' to be better than w in predicting y from x.

3.1 Hidden Attributes as Extra Information

In the object recognition with hidden attribute setting, hidden attributes can be utilized as additional information to learn a mapping function $g : \mathcal{A} \longmapsto \mathbb{R}$ with parameter w^* for the prediction of class label y. In this paper, we call the g mapping function with parameter w^* as the hypothetic model. And also, the f function $f : \mathcal{X} \longmapsto \mathbb{R}$ could be called as the target model.

As shown in Figure 2, during training, the hypothetic model and target model share the same object class labels $\{y_i\}_{i=1,\ldots,N}$, but the information input for the hypothetic model is $\{a_i\}_{i=1,\ldots,N}$ which is the ground truth attributes.

Since both the image feature x and the hidden attributes a describe the object for each training sample, our basic idea in this formulation is to link the hypothetical model and the target model by regularizing the prediction score $f(x; w)$ of

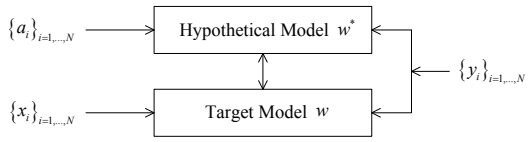

Figure 2: The hypothetical model and the target model for object recognition with hidden attributes. The hypothetical model is learned with attributes, and it helps improve the learning of the target model.

the target model to be close to the prediction score $g(a; w^*)$ of the hypothetical model. We denote this regularization as the dissimilarity regularization.

Suppose the loss functions of the target model and hypothetic model for the ith training sample to be represented by $l(y_i, x_i; w)$ and $l(y_i, a_i; w^*)$ respectively. Under the dissimilarity regularization, we would learn the parameters w and w^* of two models simultaneously with the dissimilarity regularization term, and the two parameter regularization terms $\|w\|_2^2$ and $\|w^*\|_2^2$ incorporated as shown in Equation 2:

$$\min_{w,w^*} \sum_{i=1}^{N} l(y_i, x_i; w) + \eta \sum_{i=1}^{N} l(y_i, a_i; w^*) + \frac{\lambda}{2} \sum_{i=1}^{N} \{f(x_i; w) - g(a_i; w^*)\}^2 + \frac{\gamma_1}{2}\|w\|_2^2 + \frac{\gamma_2}{2}\|w^*\|_2^2 \quad (2)$$

where λ is a positive coefficient for the dissimilarity regularization term, γ_1 and γ_2 are the positive coefficients for the corresponding parameter regularization terms, and η is the positive weight on the loss function of the hypothetic model. With this objective function, our general approach would not only minimize the loss functions for supervised learning of both models, but also minimize the dissimilarity between the predictions of two models on the original feature space and the hidden information space respectively. The regularization term effectively ties the learning of the target classifier to that of the hypothetical model such that the hidden attributes a can influence the parameters of the target model.

The general approach described in Equation 2 incorporates a squared difference term which is similar to the squared difference terms in co-regularization based multi-view semi-supervised learning (SSL) approaches [Krishnapuram *et al.*, 2005; Sindhwani and Rosenberg, 2008; Sindhwani *et al.*, 2005; Farquhar *et al.*, 2005; Belkin *et al.*, 2006], which are also regarded as the "co-training" approaches [Blum and Mitchell, 1998; Zhou and Li, 2010]. However, different from the SSL setting which focuses on better utilizing the additional unlabeled training data, the scenario of classification with hidden information assumes that hidden information is only available during training but not available during testing. We want to use such information to improve the target object classifier built on primary features. Moreover, the squared difference term in co-regularization SSL approaches are imposed on the unlabeled samples while it is imposed on labeled samples in our proposed approach.

The proposed method can be applied to different types of linear classifiers such as logistic regression (LR) and support vector machine (SVM) by selecting different loss functions. For instance, if we use hinge loss as $l(y_i, x_i; w) = \max(0, 1 - y_i w^\top x_i)$, the objective function in Equation 2 would apply our additional information modeling to the SVM learning. Since the hinge loss is still convex, subgradient based optimization can be used to solve the objective. In this paper, we apply the it to the LR model. The loss function of LR model is:

$$l(y_i, x_i; w) \triangleq -\ln p(y_i|x_i; w) = \ln\left(1 + \exp(-y_i w^\top x_i)\right)$$

where $y_i \in \{-1, 1\}$. Since the LR loss function term is convex and differentiable, gradient based methods can be applied to solve the objective function.

For linear models, the dissimilarity regularization term in Equation 2 can be written as:

$$\frac{\lambda}{2} \sum_{i=1}^{N} \{f(x_i; w) - g(a_i; w^*)\}^2 = \frac{\lambda}{2}(\mathbf{X}w - \mathbf{A}w^*)^\top(\mathbf{X}w - \mathbf{A}w^*) \triangleq \frac{\lambda}{2}\mathbf{w}^\top C\mathbf{w} \quad (3)$$

where $\mathbf{X} = [x_1, x_2, \ldots, x_N]^\top$ denotes the training data matrix, $\mathbf{A} = [a_1, a_2, \ldots, a_N]^\top$ denotes the hidden information matrix, $C = [\mathbf{X}, -\mathbf{A}]^\top[\mathbf{X}, -\mathbf{A}]$, and $\mathbf{w} = [w^\top, w^{*\top}]^\top$.

Since $\mathbf{w}^\top C\mathbf{w} = \{[\mathbf{X}, -\mathbf{A}]\mathbf{w}\}^T\{[\mathbf{X}, -\mathbf{A}]\mathbf{w}\} \geq 0$, the matrix C is positive semi-definite for any vector $\mathbf{w}$. Thus, the score similarity term is also a convex quadratic term. Its gradient with respect to vector $\mathbf{w}$ is:

$$\nabla_{\mathbf{w}} \frac{\lambda}{2} \sum_{i=1}^{N} \{f(x_i; w) - g(a_i; w^*)\}^2 = \lambda C\mathbf{w} \quad (4)$$

This gradient can be directly combined with the gradients of the remaining terms in Equation 2 to optimize the objective for the learning of parameters w and w^*.

3.2 Object-Attribute Relationships as Hidden Information

To further improve the performance of object recognition with hidden attributes, we propose to exploit the additional information in the attributes, i.e. the relationships between objects and attributes. As a set of semantic descriptions about the objects, attributes hold strong relationships with categories of objects that are determined by the intrinsic properties of different categories of objects. For instance, the object "bird" holds co-occurrence relationship with attribute "has wing", and holds mutually exclusive relationship with attribute "has horn". We believe such relationships, if captured as additional hidden information, would enforce the object classification to fit not only with the object labels, but also with the intrinsic properties of objects. In this way, the classifiers learned with attributes as hidden information can generalize better in the testing data.

To simplify the analysis, we consider the relationship between object label y and each of the M types of attributes, i.e. a^m with $m \in [1, M]$, in a pairwise manner. Suppose the relationship between object label y and attribute a^m can be evaluated by a d dimensional real valued vector t_{ym}. Also, the relationship between predicted object $\hat{y}$ and attribute a^m can be evaluated by another d dimensional real valued vector $\hat{t}_{ym}$. Since the predicted object $\hat{y}$ is given by the mapping

function $f : \mathcal{X} \longmapsto \mathbb{R}$ with parameter w, $\hat{t}_{ym}$ should also be a function of w as $\hat{t}_{ym}(w)$.

The essence of our method here is enforcing the relationship between the predicted object $\hat{y}$ and each of the attribute label to be close to the relationship between the ground truth object label and the corresponding attribute label. In this way, our classifier learning can be connected with the hidden information. Such a regularization is natural, since a perfect object classification should also preserve the relationships between objects and attributes perfectly. Suppose the above two relationships can be evaluated by vectors $\hat{t}_{ym}(w)$ and t_{ym} respectively, we can hence enforce the ℓ_2 norm of the vector difference $\hat{t}_{ym}(w) - t_{ym}$ to be small. Combining such a relationship regularization with the terms for standard object classifier learning as in Equation 1, our general formulation for exploiting object-attribute relationships can then be written as in Equation 5.

$$\min_{w} \sum_{i=1}^{N} l(y_i, x_i; w) + \frac{\gamma}{2}\|w\|_2^2 + \frac{\zeta}{2}\sum_{m=1}^{M} \|\hat{t}_{ym}(w) - t_{ym}\|_2^2 \quad (5)$$

Compared to the formulation in Equation 2, this formulation does not require to learn a hypothetic classifier and is therefore more computationally efficient.

To fulfill the general formulation in Equation 5, we further introduce the detailed definition of $\hat{t}_{ym}(w)$ and t_{ym}. We utilize the linear regression coefficients to evaluate the relationships between object y and each of the attribute a^m. Here, the regression coefficients r_{ym} and s_{ym} reconstruct the object label y from the attribute a^m as $y = r_{ym} + s_{ym}a^m$. Coefficients r_{ym} and s_{ym} can then be obtained by minimizing the mean square error as:

$$\min_{r_{ym}, s_{ym}} \frac{1}{N}\sum_{i=1}^{N}(y_i - r_{ym} - s_{ym}a_i^m)^2$$

where a_i^m is the value of attribute a^m for sample i.

Both the coefficients r_{ym} and s_{ym} have their specific meanings for representing the relationship. When y and a^m are binary values with "1" standing for positive label and "-1" standing for negative label, s_{ym} will reflect the "co-occurrence" ($s_{ym} > 0$) and "mutually exclusive" ($s_{ym} < 0$) relationships with its amplitude indicating the extent of the relationship. When $s_{ym} \approx 0$, the two variables tend to be "unrelated". Also, r_{ym} represents the bias between y and a^m values. It gives the prior information on whether y would be more frequent to present than a^m or not.

Here, we define the matrix ϕ_m, the object label vector $\mathbf{y}$, and the relationship evaluation vector t_{ym} as:

$$\phi_m = \begin{bmatrix} a_1^m & 1 \\ \vdots & \vdots \\ a_N^m & 1 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \qquad t_{ym} = \begin{bmatrix} s_{ym} \\ r_{ym} \end{bmatrix}$$

Vector t_{ym} can then have a closed form solution as $t_{ym} = (\phi_m^\top \phi_m)^{-1}\phi_m^\top \mathbf{y} \equiv \phi_m^+ \mathbf{y}$, where ϕ_m^+ is the Moore-Penrose pseudo-inverse [Penrose, 1955] of matrix ϕ_m. Given the object and attribute labels in the training data, vector t_{ym} should be a constant unrelated to classifier parameter w.

The predicted object $\hat{y}$ can be further represented by the object classifier response $w^\top x$. Hence, the regression coefficients $\hat{r}_{ym}$ and $\hat{s}_{ym}$ should reconstruct $w^\top x$ from attribute a^m as $w^\top x = \hat{r}_{ym} + \hat{s}_{ym}a^m$. These two coefficients are obtained by minimizing the following mean square error:

$$\min_{\hat{r}_{ym}, \hat{s}_{ym}} \frac{1}{N}\sum_{i=1}^{N}(w^\top x_i - \hat{r}_{ym} - \hat{s}_{ym}a_i^m)^2$$

Define the training sample matrix $\mathbf{X}$, and the relationship evaluation vector $\hat{t}_{ym}(w)$ as:

$$\mathbf{X} = \begin{bmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{bmatrix} \qquad \hat{t}_{ym}(w) = \begin{bmatrix} \hat{s}_{ym} \\ \hat{r}_{ym} \end{bmatrix}$$

then the closed form solution for relationship evaluation vector is $\hat{t}_{ym}(w) = \phi_m^+ \mathbf{X} w$.

Now, we replace the $\hat{t}_{ym}(w)$ term in Equation 5 as $\phi_m^+\mathbf{X}w$, and keep the pre-calculated constant term t_{ym}. The complete objective function formulation can then be given in Equation 6.

$$\begin{aligned} &\min_{w} \sum_{i=1}^{N} l(y_i, x_i; w) + \frac{\gamma}{2}\|w\|_2^2 \\ &+\frac{\zeta}{2}\sum_{m=1}^{M}\left\{\left(\phi_m^+\mathbf{X}w - t_{ym}\right)^\top\left(\phi_m^+\mathbf{X}w - t_{ym}\right)\right\} \end{aligned} \quad (6)$$

From Equation 6, we can see our formulation capturing object-attribute relationship would bring in an additional quadratic term $w^\top\mathbf{X}^\top(\phi_m^+)^\top\phi_m^+\mathbf{X}w$. Reshaping this term, we find it equals to $(\phi_m^+\mathbf{X}w)^\top(\phi_m^+\mathbf{X}w) \geq 0$ for any w. Hence, the matrix $\mathbf{X}^\top(\phi_m^+)^\top\phi_m^+\mathbf{X}$ is positive semi-definite, and the whole quadratic term is then convex. Such a quadratic convex term is easy to optimize.

Similar to our previous formulation discussed in Section 3.1, the formulation in Equation 6 can also be applied to different types of linear classifiers such as logistic regression (LR) and support vector machine (SVM) by selecting different loss functions. As discussed in Section 3.1, if we use hinge loss as $l(y_i, x_i; w) = \max(0, 1 - y_i w^\top x_i)$, the objective function in Equation 6 would apply our relationship modeling to the SVM learning. Since the hinge loss is still convex, subgradient based optimization can be used to solve the objective. In this paper, we apply the relationship modeling to the LR model. The optimization of the LR loss functions have been discussed in Section 3.1.

For our formulation in Equation 6, the gradient of the relationship term can be represented as:

$$\begin{aligned} &\nabla_w \frac{\zeta}{2}\sum_{m=1}^{M}\left\{\left(\phi_m^+\mathbf{X}w - t_{ym}\right)^\top\left(\phi_m^+\mathbf{X}w - t_{ym}\right)\right\} \\ =&\zeta\sum_{m=1}^{M}\left\{\mathbf{X}^\top(\phi_m^+)^\top\phi_m^+\mathbf{X}w - \mathbf{X}^\top(\phi_m^+)^\top t_{ym}\right\} \end{aligned} \quad (7)$$

Since the relationship term in Equation 6 is still convex, we can directly combine the gradient in Equation 7 with the gradients of LR loss functions and the ℓ_2 norm parameter regularization term for object classifier learning.

3.3 Combined Formulation

The formulation discussed in Section 3.1 utilizes hidden attributes as additional information, and enforce the score dissimilarity between the hypothetic model on the hidden attributes and the target model on the image feature to be small. On the other hand, the formulation in Section 3.2 models the relationships between hidden attributes and the object. It enforces preservation of the relationships between attributes and the object category. These two formulations incorporate different properties of hidden attributes, and hence can be further combined into one objective function. The combined objective function can be written as:

$$\begin{aligned}&\min_{w,w^*} \sum_{i=1}^{N} l(y_i, x_i; w) + \eta \sum_{i=1}^{N} l(y_i, a_i; w^*) \\ &+ \frac{\lambda}{2} \sum_{i=1}^{N} \{f(x_i; w) - g(a_i; w^*)\}^2 + \frac{\gamma_1}{2}\|w\|_2^2 + \frac{\gamma_2}{2}\|w^*\|_2^2 \\ &+ \frac{\zeta}{2} \sum_{m=1}^{M} \sum_{m=1}^{M} \left\{ \left(\phi_m^+ \mathbf{X} w - t_{ym}\right)^\top \left(\phi_m^+ \mathbf{X} w - t_{ym}\right) \right\} \end{aligned} \quad (8)$$

This objective function hence minimizes the score dissimilarity term and the relationship regularization term simultaneously during the learning of parameters w and w^*.

The gradient of the objective function in Equation 8 with respect to w and w^* is the combination of the gradients for each term in this equation. The gradients of our proposed score dissimilarity term and the relationship regularization term are given in Equation 4 and Equation 7 respectively. Since each term in Equation 8 is convex, the optimization can be solved by gradient descent based methods.

4 Experiments

We perform experiments on natural scene object classification on two benchmark datasets: aPascal dataset [Farhadi *et al.*, 2009] and Animals with Attributes (AWA) dataset [Lampert *et al.*, 2009]. The goal is to compare the performance of our proposed approaches incorporating hidden attributes with the basic approaches without using attributes, the traditional attribute approaches using attributes as the middle level representation, and the existing approaches for learning with hidden information.

Models. The models evaluated in our experiments include: the standard logistic regression model (**LR**) and support vector machine (**SVM**) learned with only the training data, the proposed formulation in Equation 2 using attributes as extra information (**xLR+**), the proposed formulation in Equation 6 with incorporating attribute relationships (**LR-Rel+**), the formulation in Equation 8 further combining formulations in Equation 2 and Equation 6 (**xLR-Rel+**).

The aPascal dataset contains 6340 training images and 6355 testing images collected from Pascal VOC 2008 challenge. Each sample belongs to one of the twenty object categories: *people, bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa,* and *tv/monitor*. The dataset also provides 9751 dimensional base feature for each of the training and testing sample. Various types of color, textural, HOG, shape and edge descriptors are combined with a Bag-of-Words approach to formalize the 9751 dimensional base feature vector. The base feature is used in all of the following experiments.

A list of 64 attributes are annotated for each sample in the dataset with examples shown in Figure 1. Each attribute is quantized into "-1" or "1" binary values to represent the absence or presence of the attribute. These hidden attributes are used for the proposed algorithms xLR+, LR-Rel+ and xLR-Rel+. We use a one-versus-all strategy to perform multi-class object classification in this dataset. A total of 20 classifiers are trained to predict each category against the remaining categories. The final decision is made by comparing the scores for each object type.

During classifier learning, the coefficients are tuned through a two fold cross validation procedure within the training set. The results are shown in Table 1, where both the overall accuracy rate and mean per-class recognition accuracy rate are given. We also include the results from state of art middle level representation work [Farhadi *et al.*, 2009; Wang and Mori, 2010; Wang and Ji, 2013], and the results from the state of art learning with hidden information approaches include the SVM+ approach [Vapnik and Vashist, 2009] and the LR+ approach [Wang *et al.*, 2014] in Table 1.

Table 1: Object recognition results on aPascal dataset compared to state of the art middle level representation based attribute methods.

Methods (%)	Model	Mean	Overall
	LR	43.29	59.76
This work	xLR+	43.87	60.23
This work	LR-Rel+	47.52	62.03
This work	xLR-Rel+	47.82	**63.10**
[Farhadi *et al.*, 2009]	SVM	37.70	59.40
[Wang and Mori, 2010]	LSVM	**50.84**	59.15
[Wang and Ji, 2013]	BN	44.82	63.02
[Vapnik and Vashist, 2009]	SVM+	42.08	60.02
[Wang *et al.*, 2014]	LR+	42.21	60.17

Firstly, we compare the baseline LR approach with our proposed models xLR+, LR-Rel+, and xLR-Rel+. From the results, we can see that by incorporating hidden attributes, the proposed models xLR+, LR-Rel+, and xLR-Rel+ all outperform LR in terms of both the overall and mean per-class accuracies, which shows the effectiveness of the proposed algorithms. In addition, xLR-Rel+ outperforms both xLR+ and LR-Rel+ in both accuracy evaluations. This comparison shows that the combination further improves the performance.

Secondly, we compare with state of art middle level representation approaches by [Farhadi *et al.*, 2009], [Wang and Mori, 2010] and [Wang and Ji, 2013]. We can see that although predicted attributes are not utilized during testing, all three of our models (i.e. xLR+, LR-Rel+ and xLR-Rel+) outperform the approach proposed in [Farhadi *et al.*, 2009]. Compared to the results by [Wang and Mori, 2010], our xLR-Rel+ approach performs better in overall recognition rate by

around 4%, and performs lower in mean per-class recognition rate by about 3%. This is expected. As argued in [Wang and Ji, 2013], [Wang and Mori, 2010] use the loss function specifically designed for skewed data, and the aPascal data is skewed by having 2571 of 6355 testing samples to be in "person" category. [Wang and Mori, 2010] also report their performances with the standard "0/1" loss function. Results are 46.25% for mean accuracy, and 62.16% for overall accuracy, which are both not as good as our performances. The approach in [Wang and Ji, 2013] also combines the attribute relationship in the model, and its object recognition performance in aPascal is not as good as our xLR-Rel+ model for both overall and mean per-class evaluations. These results show that our approaches are quite effective for improving object classifier learning compared to traditional middle level representation methods.

Thirdly, we compare our methods with learning with hidden information approaches including the SVM+ approach [Vapnik and Vashist, 2009] and the LR+ approach [Wang *et al.*, 2014]. From Table 1, we can see all our three models (i.e. xLR+, LR-Rel+ and xLR-Rel+) outperforms both SVM+ and LR+ approaches in both the mean and overall recognition accuracy. We perform the Wilcoxon rank sum test to evaluate the performance improvement of the proposed xLR-Rel+ model over both LR and SVM+ approaches. Both tests show performance improvements are statistically significant with a p-value less than 5%.

To further compare the performances of our proposed model with the learning with hidden information state of the arts including SVM+ and Rank Transfer, we test the proposed algorithm for object classification on the Animals with Attributes (AWA) dataset [Lampert *et al.*, 2009]. This dataset includes 6180 images that belong to 10 testing classes. These 10 testing classes are different wild animals including *chimpanzee* (CP), *giant panda* (GP), *leopard* (LP), *persian cat* (PC), *pig* (PG), *hippopotamus* (HP), *humpback whale* (HW), *raccoon* (RC), *rat* (RT), and *seal* (SL).

To compare with the results in [Sharmanska *et al.*, 2013], we follow the same experimental setting as in [Sharmanska *et al.*, 2013]. In such setting, the models are tested on recognizing each possible pair of the 10 animal classes. This would give us 45 animal pairs. Also, the provided SURF descriptor in 2000 dimensions are used as features, and the predicted attributes in the format of probability estimates provided by [Lampert *et al.*, 2009] are used as hidden information during training. With the provided feature and hidden information, 45 binary object classifiers are trained for each animal pair. We use 100 samples per object class for training, and 200 samples per object class for testing. As in [Sharmanska *et al.*, 2013], we repeat such training/testing split procedure for 20 times. The results with comparisons to SVM+ and Rank Transfer methods are presented in Table 2

From the results in Table 2, our proposed xLR-Rel+ is very effective for incorporating hidden attributes. Among the 45 possible cases, the SVM performs the best only in 1 case, the SVM+ performs the best in 7 cases, the Rank Transfer method performs the best in 11 cases, and our proposed xLR-Rel+ model performs the best in 26 out of the 45 cases. The average values of accuracies over the total 45 pairs also show that our proposed model performs better than SVM, SVM+, and Rank Transfer methods.

Table 2: Object Recognition with Hidden Attributes on AWA Dataset

		SVM	SVM+	Rank Transfer	xLR-Rel+
1	CP vs. GP	91.53	**92.12**	91.83	83.85 ± 1.45
2	CP vs. LP	94.16	94.23	94.80	**98.03 ± 1.03**
3	CP vs. PC	91.09	91.73	91.86	**95.17 ± 1.24**
4	CP vs. PG	87.45	88.06	**88.59**	86.57 ± 1.58
5	CP vs. HP	**87.58**	87.53	87.57	87.08 ± 1.65
6	CP vs. HW	98.12	98.57	98.52	**99.60 ± 0.88**
7	CP vs. RC	89.00	**89.67**	89.54	87.98±1.43
8	CP vs. RT	86.84	87.96	88.47	**92.95 ± 2.15**
9	CP vs. SL	92.53	**92.59**	92.58	90.54 ± 2.19
10	GP vs. LP	95.13	94.95	95.11	**97.74 ± 0.92**
11	GP vs. PC	94.66	**94.68**	94.38	93.27 ± 1.86
12	GP vs. PG	88.67	**88.95**	88.69	81.87 ± 1.70
13	GP vs. HP	92.35	**92.85**	92.78	88.93 ± 1.68
14	GP vs. HW	98.77	98.76	98.88	**98.97 ± 0.73**
15	GP vs. RC	91.76	**91.90**	91.33	86.99 ± 1.91
16	GP vs. RT	90.50	90.61	90.33	**90.69 ± 1.15**
17	GP vs. SL	93.33	93.40	**93.58**	89.85 ±1.05
18	LP vs. PC	95.50	95.65	95.92	**97.65 ± 1.11**
19	LP vs. PG	90.40	90.40	90.88	**96.95 ± 0.87**
20	LP vs. HP	93.60	93.83	93.81	**96.12 ± 1.30**
21	LP vs. HW	99.06	99.20	99.17	**99.43 ± 1.41**
22	LP vs. RC	83.23	83.18	83.15	**90.66 ± 2.84**
23	LP vs. RT	90.28	90.65	90.98	**96.50 ± 1.44**
24	LP vs. SL	94.98	95.14	95.49	**97.09 ± 1.59**
25	PC vs. PG	83.23	83.38	**83.39**	78.31 ± 1.99
26	PC vs. HP	92.66	93.14	93.41	**94.14 ± 0.93**
27	PC vs. HW	96.19	96.69	97.26	**99.64 ± 1.42**
28	PC vs. RC	90.46	90.94	**91.20**	88.40 ± 1.67
29	PC vs. RT	69.38	69.43	**70.40**	68.41 ± 1.89
30	PC vs. SL	86.06	86.97	86.91	**90.43 ± 1.78**
31	PG vs. HP	76.45	77.42	79.02	**82.01 ± 2.72**
32	PG vs. HW	96.78	97.04	97.32	**98.66 ± 1.49**
33	PG vs. RC	80.08	81.50	**81.79**	78.13 ± 2.10
34	PG vs. RT	72.25	72.63	73.68	**73.70 ± 2.90**
35	PG vs. SL	79.76	80.33	**81.76**	78.32 ± 2.35
36	HP vs. HW	93.83	93.63	93.75	**98.17 ± 1.42**
37	HP vs. RC	86.49	86.83	**87.37**	84.21 ± 1.80
38	HP vs. RT	85.12	85.99	87.37	**90.55 ± 1.89**
39	HP vs. SL	72.82	73.41	**75.85**	70.98 ± 3.16
40	HW vs. RC	96.92	97.11	97.15	**99.32 ± 1.12**
41	HW vs. RT	95.21	95.45	95.53	**99.39 ± 0.92**
42	HW vs. SL	86.44	86.89	86.93	**96.80 ± 2.86**
43	RC vs. RT	79.59	79.67	**80.31**	79.49 ± 2.71
44	RC vs. SL	92.22	92.55	**92.80**	83.09 ± 1.37
45	RT vs. SL	80.44	80.68	82.34	**88.02 ± 2.55**
*	*Average*	88.95	89.30	89.64	**89.88**

5 Conclusion

In this work, we propose to incorporate hidden attributes for object classification. Instead of predicting these attributes explicitly or implicitly during testing, we utilize the attributes only during training to improve the learning of the object classifier on the primary features. We develop two different approaches to incorporate the hidden attributes, with one approach utilizing attributes as additional information, and the other incorporating the relationship between attributes and objects. Finally, these two different approaches are combined into one learning objective. We evaluate our approach on the natural scene object classification. Experiments demonstrate the effectiveness of our approaches for classification over state of the art methods on benchmark datasets.

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, and by the Army Research Office under grant W911NF-13-1-0395.

References

[Akata *et al.*, 2013] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, Cordelia Schmid, et al. Label-embedding for attribute-based classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 819–826, 2013.

[Belkin *et al.*, 2006] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[Blum and Mitchell, 1998] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Annual Conference on Computational Learning Theory (COLT)*, pages 92–100, 1998.

[Farhadi *et al.*, 2009] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1785, 2009.

[Farquhar *et al.*, 2005] Jason Farquhar, David Hardoon, Hongying Meng, John Shawe-Taylor, and Sandor Szedmak. Two view learning: Svm-2k, theory and practice. In *Advances in Neural Information Processing Systems (NIPS)*, pages 355–362, 2005.

[Ferrari and Zisserman, 2007] Vittorio Ferrari and Andrew Zisserman. Learning visual attributes. In *Advances in Neural Information Processing Systems (NIPS)*, pages 433–440, 2007.

[Hwang *et al.*, 2011] Sung Ju Hwang, Fei Sha, and K. Grauman. Sharing features between objects and their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1761–1768, 2011.

[Kovashka *et al.*, 2011] A. Kovashka, S. Vijayanarasimhan, and K. Grauman. Actively selecting annotations among objects and attributes. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1403–1410, 2011.

[Krishnapuram *et al.*, 2005] Balaji Krishnapuram, David Williams, Ya Xue, Alexander Hartemink, Lawrence Carin, and Mario Figueiredo. On semi-supervised classification. *Advances in Neural Information Processing Systems (NIPS)*, 17:721–728, 2005.

[Lampert *et al.*, 2009] C.H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 951–958, 2009.

[Parikh and Grauman, 2011a] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1681–1688, 2011.

[Parikh and Grauman, 2011b] D. Parikh and K. Grauman. Relative attributes. In *IEEE International Conference on Computer Vision (ICCV)*, pages 503–510, 2011.

[Penrose, 1955] Roger Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(03):406–413, 1955.

[Sharmanska *et al.*, 2013] V. Sharmanska, N. Quadrianto, and C.H. Lampert. Learning to rank using privileged information. In *IEEE International Conference on Computer Vision (ICCV)*, pages 825–832, Dec 2013.

[Sindhwani and Rosenberg, 2008] Vikas Sindhwani and David S Rosenberg. An rkhs for multi-view learning and manifold co-regularization. In *International Conference on Machine learning (ICML)*, pages 976–983, 2008.

[Sindhwani *et al.*, 2005] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML Workshop on Learning with Multiple Views*, pages 74–79, 2005.

[Vapnik and Vashist, 2009] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009.

[Wang and Ji, 2012] Xiaoyang Wang and Qiang Ji. A novel probabilistic approach utilizing clip attributes as hidden knowledge for event recognition. In *International Conference on Pattern Recognition (ICPR)*, 2012.

[Wang and Ji, 2013] Xiaoyang Wang and Qiang Ji. A unified probabilistic approach modeling relationships between attributes and objects. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2120–2127, 2013.

[Wang and Ji, 2014] Xiaoyang Wang and Qiang Ji. Attribute augmentation with sparse coding. In *International Conference on Pattern Recognition (ICPR)*, pages 4352–4357, 2014.

[Wang and Mori, 2010] Yang Wang and Greg Mori. A discriminative latent model of object classes and attributes. In *European Conference on Computer Vision (ECCV)*, volume 6315, pages 155–168, 2010.

[Wang *et al.*, 2014] Ziheng Wang, Xiaoyang Wang, and Qiang Ji. Learning with hidden information. In *International Conference on Pattern Recognition (ICPR)*, pages 238–243, 2014.

[Zhao *et al.*, 2015] Yue Zhao, Nan Zhou, Libing Zhang, Licheng Wu, Rui Zheng, Xiaoyang Wang, and Qiang Ji. Shared speech attribute augmentation for english-tibetan cross-language phone recognition. In *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 539–543, 2015.

[Zhou and Li, 2010] Zhi-Hua Zhou and Ming Li. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 24(3):415–439, 2010.