LEAF: Latent Extended Attribute Features Discovery for Visual Classification

Hua Zhang[1], Rui Wang[1], Changqing Zhang[2], Xiaochun Cao[1*]
[1]State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Computer Science and Technology, Tianjin University
zhanghua,wangrui,caoxiaochun@iie.ac.cn,zhangchangqing@tju.edu.cn

ABSTRACT

To improve the discrimination of attribute representation, in this paper, we propose to extend the traditional attribute representations via embedding the latent high-order structure between attributes. Specifically, our aim is to construct the Latent Extended Attribute Features (**LEAF**) for visual classification. Since there only exist weak label for each attribute, we firstly propose a feature selection method to explore the common feature structures across categories. After that, the attribute classifiers are trained based on the selected features. Then, the category specific graph is introduced, which is composed of single attributes and their co-occurrence attribute pairs. This attribute graph is used as the initialized representation of each image. Considering our aim, we should discover the discriminative latent structure between attributes and train the robust category classifiers. To that end, we develop a joint learning objective function which is composed of the high-order representation mining term and the classifier training term. The mining term can both preserve category-specific information and discover the common structure between categories. Based on the discovery representation, the robust visual classifiers could be trained by the classifier term. Finally, an alternating optimization method is designed to seek the optimal solution of our objective function. Experimental results on the challenging datasets demonstrate the advantages of our proposed model over existing work.

CCS CONCEPTS

•Computing methodologies →Artificial intelligence; Computer vision representations; Hierarchical representations;

KEYWORDS

High-order attribute correlation discovery; semantic feature representation; visual classification

1 INTRODUCTION

Visual attribute [3, 9, 22, 26, 30, 36, 38] plays an important role in the computer vision community for describing the specific semantic visual content of images, especially for describing the common visual patterns between different categories. Furthermore, attribute serving as the image descriptor has shown impressive preliminary results on several computer vision tasks, such as, zero-shot learning [16, 20], object classification and recognition [14, 24, 30, 37], and image retrieval based on multi-attribute queries [4, 7, 28, 38, 40].

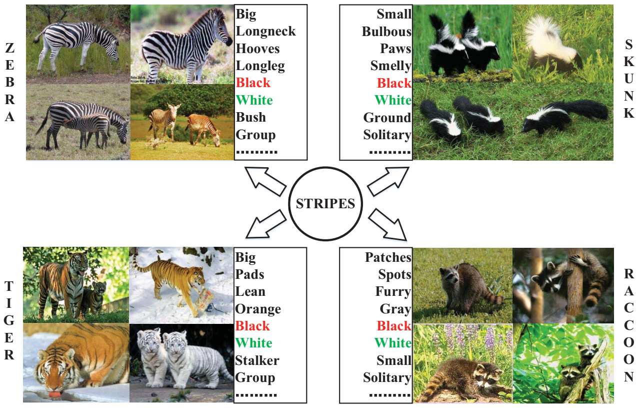
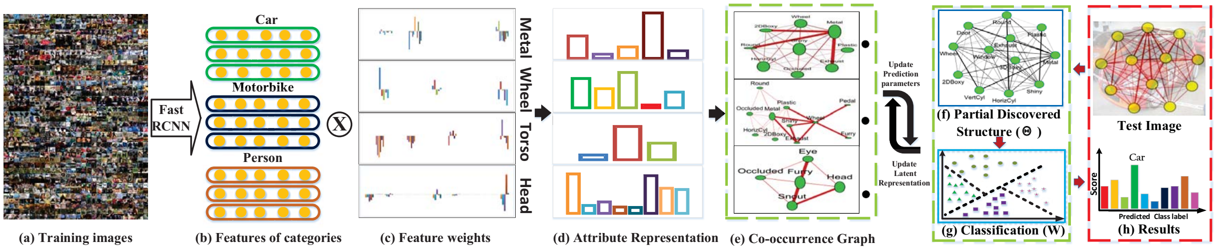


Figure 1: Illustration of training images for the attribute "Stripes" and co-occurrence attributes of different categories. We observe that different categories show distinct visual patterns for the same attribute. Furthermore, not all the co-occurrence attributes would help generate discriminative feature representation, e.g. "Black" and "White". While the high order correlations express discriminative ability, such as "Stripe-Orange-Big-Pads" for tiger.

To construct the image representation based on attributes, attribute classifiers is usually need to be learned in advance. Generally, traditional methods [27, 29, 33, 34, 39] treat the attribute learning as a weakly supervised problem, which follows the similar pipeline: Firstly, they extract the global feature of training images, and then train a group of attribute classifiers, one for each attribute. The prediction score of these classifiers are used to represent images. Finally, the classifier of each category is trained based on attribute feature representations. One of the underlying assumption of the existing methods [2, 9, 19, 20] is that distinct objects should have the consistent appearance for their shared attributes. This assumption makes the attribute ambiguous, which means one attribute refers to several different instance patterns as shown Fig. 1. For example, "Stripes" with zebra, tiger, skunk and raccoon. Thus, the discrimination of attribute as the feature representation is restricted, especially for visual classification.

[*] corresponding author

MM'17, October 23–27, 2017, Mountain View, CA, USA.
© 2017 ACM. ISBN 978-1-4503-4906-2/17/10…$15.00
DOI: http://dx.doi.org/10.1145/3123266.3123342



(a) Training images (b) Features of categories (c) Feature weights (d) Attribute Representation (e) Co-occurrence Graph (f) Partial Discovered Structure (Θ) (g) Classification (W) (h) Results

Figure 2: The flowchart of our proposed method. The training and testing steps are shown in green and red bounding box, respectively. (a) Given a set of training images, (b) we firstly extract the feature representation of each category. Then, (c) the common features of each attribute is explored to train the robust attribute classifiers. (d) Based on the prediction score of these classifiers, we obtain the attribute representation. After that, (e) we construct the co-occurrence context embedding representation by introducing the co-occurrence graph of each category. A unified discriminative pattern mining framework is proposed to discover the discriminative visual pattern (f) and prediction model parameters (g). In testing phase following with the three red arrows, we first develop the novel representation by embedding the discovered discriminative structure, and then the test image is classified based on the training models.

As suggested in [6], one straight way to relax the assumption is to train the category-specific/object-specific attribute classifiers. However, if we directly train the category-specific attribute, this would lose the semantic property of attributes sharing with different categories. Furthermore, the performance of the specific attribute classifier is dependent on the number of training images.

It has been discussed in [33, 34] that the relationship between attributes plays an important roles to improve the discrimination of attribute representation. For example, "Furry" combined with "Legs" may indicate the "Dog", while with "Cotton" would imply "Towel". Thus, the correlation between attributes is essential for attribute based visual application. Existing work [33, 34] only embed the co-occurrence between attributes into the feature representation. While such correlation between attributes may not always be benefit for obtaining the discriminative representation, especially when the attributes have the high correlations. As shown in Fig. 1, the attribute "Black" and "White" are co-occurring with "Stripes" in all the four categories. Further, All the existing work do not consider the high-order correlations between attributes. As shown in Fig. 1, the combination "Stripe-Orange-Big-Pads" have encoded the discriminative information for tiger.

To address above mentioned problems, in this paper, we introduce a novel approach to extend the traditional attribute representation via embedding the discriminative high-order correlation of attributes. Our **L**atent **E**xtended **A**ttribute **F**eatures (**LEAF**) is achieved by discovering the latent discriminative structure in attribute representation. We first extract the object proposals from images using Selective Search [31], and then construct the feature representation using Fast RCNN [15] or traditional feature detectors as shown in Fig. 2 (b). Next, we cluster these proposals based on their feature similarity by K-means and use the centers as the image feature representations. Then, a feature selection method is proposed to choose the common features sharing across categories as shown in Fig. 2 (c). Secondly, we train the attribute classifiers based on the selected features and describe the training images by the predication scores of classifiers as shown in Fig. 2 (d). The co-occurrence graph of each class is constructed based on the attribute co-occurrence matrix by computing the conditional probability of each attribute pair as shown in Fig. 2 (e). Afterwards, all the class graph are concatenated to develop our proposed co-occurrence graph representation. To discover the discriminative representation for different categories, we introduce the latent representation, which can encode both the relationships between attribute pairs and the high order correlations. Specifically, we propose a joint learning method which can discover the discriminative attribute patterns (Fig. 2 (f)) and train the classifiers (Fig. 2 (g)). The objective function is formulated as a joint optimization problem over both the latent representations and the classifier parameters. Finally, we develop an algorithm employing projected gradient descent to obtain the optimal solution. Our algorithm solves the learning problem through the alternating optimization steps between dealing with discovering the discriminative latent representation for each category and the classification parameters iteratively. During testing, the graph representation of the test image can directly combine with the discovered latent attribute patterns and classification parameters to get the prediction label (Fig. 2 (h)). To validate the effectiveness and efficiency of our proposed models, we conduct the experiments on the standard datasets for the task of image and video classification. The experimental results show a significant improvement comparing with several baselines and state-of-the-art methods, which further demonstrates the advantages of our model.

In summary, the contributions of our proposed method can be listed in three folds: 1) We propose a novel feature representation named LEAF to describe image for visual classification. Different from the traditional methods to discover the discriminative representation based on attributes and their pairwise relevance, our proposed model is composed of four layers (image, attributes, latent attribute patterns, and categories) which can encode the higher order correlations between attributes by latent layers. 2) Different from the traditional methods training the category-independent attribute classifiers, we introduce to train the robust category-specific attribute classifiers. To this end, we propose a multi-task learning framework to discover the common feature space between different categories. In this way, we not only can train the category-specific attribute classifier, but also the attribute classifiers are preserving the semantic consistency among categories. 3) To encode the high-order correlations between attribute into feature representation, we introduce a unified objective function including feature discovery and classification error term with the fisher regularization, which use the mined discriminative representation to train the classifiers.

2 RELATED WORK

In recent years, attribute [11] is widely used in the computer vision field, like object describing [9], zero-shot learning [20], object recognition [12, 23, 33, 34], image search and retrieval [25, 28], and action/event recognition [1, 13]. According to different applications, each method focuses on the distinct views to model the attributes, objects and their associations. Farhadi *et al.* [9] first propose that the images can be described by the predefined attribute and demonstrated the effectiveness of attributes on describing the images. Farhadi *et al.* [9] aim to train the discriminative attribute classifiers by feature selection. Under the task of image classification, the majority of existing methods focus on computing the correlation between attributes and the object labels. While in [6], the authors propose to train the category-specific attribute classifiers via a tensor decomposition method. Although some expected results are achieved on attribute predictions, it does not consider the correlation between attributes, which would degrade its discrimination.

Different from [9], Wang *et al.* [34] develop a model for image classification by exploiting the correlation between attributes and object labels. In [34], the attributes are firstly seen as the latent variables, and then the latent SVM [10] is introduced to compute the differences between the object labels. Although [34] obtains some promising results on image classification, all the attributes are assumed to contribute equally for all the classes. To relax this constraint, Wang *et al.* [33] develop a unified probabilistic framework to model the correlation between attributes and the class. Unlike [34] which directly constructs the correlation between attributes and object from the groundtruth label, Wang *et al.* [33] propose to explore the correlation by using the Bayesian network [21] which could give the attribute different weighting based on its importance to that object. Different from aforementioned work, our proposed method could explore the discriminative configuration based on different tasks and the discovered structures are embedded into the feature representations. The main difference between our proposed method and [33, 34] is that we jointly learn the discriminative representation and train the classifiers. Our proposed method can explore the high-order correlations between attributes.

There are also several work [1, 39] which are devoted to exploring the discriminative attribute pairs for each class. Specifically, Yuan *et al.* [39] propose to mine the discriminative co-occurrence patterns for each category. To that end, the AND/OR model is proposed to obtain the conjunction (AND) and disjunction (OR) of binary attribute features. In [1], the authors propose a semantic based video classification method which utilizes the co-occurrence semantic features as the context representation. The authors transfer the explore discriminative semantic pair problem into Generalized Maximum Clique Problem (GMCP). Specifically, it first segments videos into several clips, and then computes the attribute score of each clips. Afterwards, an all-connected graph is constructed over all the clips. Finally, the strongest clique of co-occurrence concepts are merged into the feature representation which yields a richer representation. The difference between our method and the aforementioned work is that our proposed latent representation is encoded both the relationship of the attribute pair and the high order correlation between attributes. Moreover, our proposed classifier is more robust to the uneven distribution of training images.

Recently, there exist an amount of work for visual classification via convolutional neural network (CNN) [17], and achieve some expecting results. Ross et al. [15] propose a novel framework named R-CNN for image classification and object detection. In this work, they first extract the proposals, and then employ the convolutional neural network to compute the features. To improve the efficient of feature extraction, they also propose Fast RCNN for image classification. Different from their work, our goal is to construct the attribute feature representation instead of non-semantic features.

3 PROPOSED APPROACH

Before detailing our method, we first define the notations used in this paper. A training sample is denoted as a tuple $(P_i, \mathbf{a}_i, \mathbf{y}_i)$, where P_i is the i^{th} training images. The object label of the image is represented by $\mathbf{y}_i \in \{0, 1\}^{C\times 1}$, where C is the number of classes. The attributes of image P_i are represented by a K-dimensional vector $\mathbf{a}_i = \{a_1, ..., a_k, ..., a_K\} \in \mathbb{R}^{K\times 1}$, where a_k is the indicator corresponding to the k^{th} attribute of the image. Here we assume that all the attribute classifiers are pre-trained. We attempt to firstly introduce the method to train the category-specific attribute classifiers. Then, the co-occurrence graph representation is constructed based on the attribute representations. Finally, we propose the unified framework to simultaneously discovery the discriminative latent attribute patterns and learn the prediction parameters.

3.1 Attribute representation construction

Similar to [15], we firstly extract the object candidates from the training images, and then train the category-specific attribute classifiers based on these candidates. To extract the object candidates, we employ Selective Search [31] which is widely used in object detection and classification. Moreover, we set the threshold between any proposals is 0.3, which can suppress the cluster background. After that, there exist about 500 proposals for each image. Finally, we extract the features from these extracted proposals. Specifically, Fast RCNN [15] with pre-trained model is used to extract the feature representation of FC7 layer, whose dimension is $4,096$.

When object proposals are extracted from images, we observe that the spatial distribution of proposals are non-uniform, and some parts may relate with the majority of candidates. To avoid the attribute classifiers biased by these candidates, we cluster these proposals into K centers by K-means based on the extracted features, and then these centers are used as the feature representation.

Before directly using these representation to train the attribute classifiers, we propose a multi-task learning method to discover the latent common structure for one attribute between different categories. The assumption to learn such latent structure is that not all the dimensions of feature representation are related with one specific attribute. And one attribute may only involve a set of dimensions, sharing with different categories which includes the attribute. To that end, we propose to use a multi-task learning framework to find the latent structure for each attribute. Since we consider the sharing attribute between categories could be represented by their common feature structures, we use $y_i^{a_k} \in \{0,1\}^C$ denoting the label of i_{th} image including the k_{th} attribute. And $fc_i \in \mathbb{R}^{4096\times 1}$ is the low-level feature representation.

$$W_s = \arg min \frac{1}{2} \sum_{i=1}^{N_a} \sum_{k=1}^{C_a} ||W_s * fc_i - y_i^{a_k}||^2 + \lambda_t ||W_s||_{2,1}, \qquad (1)$$

where W_s denotes the discovered latent feature structure for the a_k attribute. λ_t is the weighting parameter for sparsity. N_a denotes the number of training images for the attribute a_k and C_a is the amount of categories containing the attribute a_k. For each attribute, we learn different latent structure for attribute classifiers. After that, we train the attribute classifiers with libsvm [5] with linear kernel based on the discovered feature representation.

3.2 Co-occurrence attribute graph representation

We denote the groundtruth attribute label of each image represented as $\mathbf{l} = \{l_{a_1}, ..., l_{a_k}\} \in \{0,1\}^{K\times 1}$. There are several attributes associated with each image. We construct the class-specific co-occurrence matrix $\mathbf{H}_c \in \mathbb{R}^{K\times K}$ by computing the frequency of training images including the specified pair of attributes as follows:

$$\mathbf{H}_c(l^c_{a_u}, l^c_{a_v}) = p(l^c_{a_u} | l^c_{a_v}) = \frac{\#(l^c_{a_u}, l^c_{a_v})}{\#(l_{a_v})}, \qquad (2)$$

where $\#(l^c_{a_u}, l^c_{a_v})$ refers to the number of training images in which both attribute $l^c_{a_u}$ and $l^c_{a_v}$ occur, and $\#(l_{a_v})$ is the amount of training images including attribute l_{a_v}. The reasons to construct the co-occurrence matrix based on the conditional probability are that: Firstly, it would release the effect caused by the limited number of training samples. Secondly, the asymmetry of $\mathbf{H}_c$ could make the class-specific attributes achieve a high weighting. This can give the high frequency combination of attributes more attention.

When the co-occurrence matrix of each class is obtained, we use a method similar with [1] to construct the co-occurrence graph for the c^{th} class, say $G_c = \{\mathcal{V}_c, \mathcal{E}_c, \mathcal{W}_c\}$ where $\mathcal{V}_c$ represents the set of nodes, the number of which is equal to that of the predefined attributes. $\mathcal{E}_c$ is the set of edges between correlated attributes that are determined by the nonzero elements in $\mathbf{H_c}$. $\mathcal{W}_c \in \mathbb{R}^{K\times K}$ means the values of edge weights by embedding the score of attributes. And the weight of each edge $\mathcal{W}_c(l^c_{a_u}, l^c_{a_v})$ is defined based on the conditional probability as:

$$\mathcal{W}_c(l^c_{a^c_u, a_v}) = \mathbf{H}_c(l^c_{a_u}, l^c_{a_v}) \times \phi(a_v), \qquad (3)$$

where $\phi(\cdot)$ is the sigmoid function which maps the attribute prediction scores between $[0,1]$ and $\mathcal{W}_c(l^c_{a_u}, l^c_{a_v}))$ could be seen as the probability of the attribute pair $(l^c_{a_u}, l^c_{a_v})$ occurring given the attribute $l^c_{a_v}$.

After constructing the co-occurrence attribute graph G_c for the c^{th} category, we combine the attribute representation $\mathbf{a}_i$ of P_i to get the graph representation $\mathbf{X}_i^c = \mathcal{W}_c \times \mathbf{a}_i \in \mathbb{R}^{K\times K}$. Since the graph representation $\mathbf{X}_i^c$ of the c^{th} category is embedded the co-occurrence of attributes, each element of the representation has the semantic information. The diagonal elements of $\mathbf{X}_i^c$ refer to the importance of each single visual attribute while the non-diagonal elements indicate to the importance of pairwise attributes. We concatenate all the class-specific graph together to describe each image. Now, the i^{th} training image can be described by $\mathbf{X}_i = [\mathbf{X}_i^1, ..., \mathbf{X}_i^C] \in \mathbb{R}^{(K\times K)\times C}$, which embeds the attribute scores and their co-occurrence correlation under different category graphs. In fact, comparing with the original feature representation only using the confidence values of attribute classifiers, our proposed graph representation is of more generality by encoding the class-specific level similarity into the feature representation *e.g.* the shared attributes between "Cow" and "Horse".

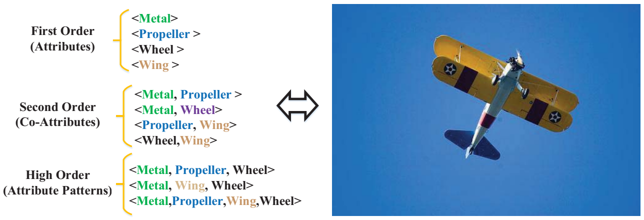


Figure 3: Illustration of relationship between attributes. Given an input image (right) in aPascal dataset and its attribute labels, we define the correlation between attributes (left) from first-order to high order.

Here, the co-occurrence matrix constructed in this section is based on the image-level attribute labels. However, this limitation can be easily relaxed to the category-level attribute annotation by assuming the images of each category labeled with the same attributes as discussed in [20]. Another limitation of class-specific attribute graph representation may be the dimension of $\mathbf{X}_i$. Furthermore, we observe that $\mathbf{X}_i$ is very sparse because each attribute only relates with several categories and each category associates with a small number of attributes. To prevent generating the high dimension description vectors, we only select the class correlated attribute as the node to construct the graph for each category. For example, on aPascal, if we averagely use 12 attributes to represent each category, which makes the graph representation very sparse. For each class, we only select these class-related attributes as the graph node. Therefore, the dimensions of feature vectors would be significantly reduced, which is about $12 \times 12 \times C$ for each image, where $C = 20$ is the number of classes.

3.3 Model learning

In this section, we expect to find the discriminative latent attribute patterns of each class based on $\mathbf{X}_i$. It is a difficult to directly extract the hidden discriminative representation from $\mathbf{X}_i$, due to the similar representations of different classes like "Dog" vs "Cat". To solve this problem, we introduce a latent representation based method to discover the discriminative representation for each category.

Let $\mathbf{z}_i \in \mathbb{R}^{M\times 1}$ be the latent attribute representation for the image P_i, where $M = K' \times K' \times C$, where K' refers the number of related attributes for current category. $\mathbf{z}_i$ is composed of the discriminative visual patterns, and the values of which derived from $\mathbf{X}_i$ contains attribute scores and the co-occurrence probability. Thus, each element of $\mathbf{z}_i$ should be non-negative. Our task is that given the graph representation $\mathbf{X}_i$ and the class labels of each image $\mathbf{y}_i$, our objective is to explore a latent attribute representation of each class for improving the performance of image classification. To this end, the objective function can be formulated as:

$$\underset{\{\mathbf{z}_i\},\Theta,\mathbf{W}}{\operatorname{argmin}} \sum_{i=1}^{N}(\alpha_i\mathcal{L}(\mathbf{z}_i, f(\mathbf{X}_i;\Theta)) + \Omega(\mathbf{y}_i, \Psi(\mathbf{z}_i;\mathbf{W}))) \qquad (4)$$
$$+\beta R(\Theta) + \gamma R(\mathbf{W})$$
$$s.t.\ \mathbf{z}^i \geq 0,\ \forall i = 1, 2, ..., N,$$

where $\mathcal{L}(\mathbf{z}_i, f(\mathbf{X}_i;\Theta))$ denotes the loss function from the input graph representation $\mathbf{X}_i$ to the discovered discriminative latent attribute patterns $\mathbf{z}_i$. $f(\cdot)$ is the regression function for mapping the graph representation to the discriminative patterns with parameter $\Theta \in \mathbb{R}^{M\times M}$. $\Omega(\mathbf{y}_i, \Psi(\mathbf{z}_i;\mathbf{W}))$ represents the misclassification errors under the latent representations $\mathbf{z}_i$, and $\mathbf{W} \in \mathbb{R}^{M\times C}$ is the prediction function which maps the latent representation to the prediction labels. $R(\cdot)$ is the regularization function on the corresponding variables to prevent the over-fitting. N is the number of training images. Moreover, α_i, β, γ are the weighting parameters corresponding to each term. And $\alpha_i = 1$ is fixed for all the samples.

Specifically, to make the objective function Eq. 4 be solved efficiently and easily, $\mathcal{L}(\mathbf{z}_i, f(\mathbf{X}_i;\Theta)$ is set to be the least square loss functions and $\Omega(\mathbf{y}_i, \Omega(\mathbf{z}_i;\mathbf{W}))$ can be seen as a linear classification loss function. Thus, each term is formulated as:

$$\mathcal{L}(\mathbf{z}_i, f(\mathbf{X}_i;\Theta) = ||\mathbf{z}_i - \Theta\varphi(\mathbf{X}_i)||_F^2, \qquad (5)$$

$$\Omega(\mathbf{y}_i, \Psi(\mathbf{z}_i;\mathbf{W}))) = ||\mathbf{y}_i - \mathbf{W}^T\mathbf{z}_i - \mathbf{b}||_2^2, \qquad (6)$$

$$\underset{\{\mathbf{z}_i\},\Theta,\mathbf{W}}{\operatorname{argmin}} \sum_{i=1}^{N}(\alpha_i||\mathbf{z}_i - \Theta\varphi(\mathbf{X}_i)||_F^2) + ||\mathbf{y}_i - \mathbf{W}^T\mathbf{z}_i - \mathbf{b}||_2^2) \qquad (7)$$
$$+\beta||\Theta||_F^2 + \gamma||\mathbf{W}||_F^2$$

where $\varphi(\mathbf{X}_i) \in \mathbb{R}^{M\times 1}$ denotes vectorization of the input matrix. $||\cdot||_F$ represents the Frobenius norm, and $||\cdot||_2$ is the Euclidean norm. Considering directly to solve the above function may easily overfit, we add the regularization term on each latent variables. Eq. 5 can be seen as a reconstruction problem, a Frobenius norm based regularization $||\Theta||_F^2$ is added to prevent from a trivial solution. To guarantee the generalization of classification model, Frobenius norm based regularization is also introduced on the classification parameters $\mathbf{W}$. Furthermore, we also enforce a ℓ_1 norm regularization function over $\mathbf{z}_i$ to promote its sparsity. Thus, we merge Eq. 5 and 6 and their corresponding regularization term into the main objective function:

$$\underset{\{\mathbf{z}_i\},\Theta,\mathbf{W}}{\operatorname{argmin}} \sum_{i=1}^{N}(\alpha_i||\mathbf{z}_i - \Theta\varphi(\mathbf{X}_i)||_F^2) + ||\mathbf{y}_i - \mathbf{W}^T\mathbf{z}_i - \mathbf{b}||_2^2) \qquad (8)$$
$$+\beta||\Theta||_F^2 + \gamma||\mathbf{W}||_F^2 + \eta\sum_i||\mathbf{z}_i||_1$$
$$s.t.\ \mathbf{z}^i \geq 0,\ \forall i = 1, 2, ..., N,$$

where η is the trade-off parameter to control the sparsity of $\mathbf{z}_i$. From the objective function Eq. 8, we could find that our proposed model can automatically discover the discriminative configurations $\mathbf{z}_i$ by keeping a small classification loss. Since we have embedded the co-occurrence of attributes into the feature representation $\mathbf{X}_i$, the latent variable $\mathbf{z}_i$ contains both the class-specific attributes and the correlations of dependent attribute pairs for our task. The latent variable $\mathbf{z}_i$ contains two parts: One is the score of each attribute, and the other is the correlation between any attributes. This makes $\mathbf{z}_i$ very sparse for every class only include the small number of dictionaries. Thus we add a sparse regularization.

The latent variables $\mathbf{z}_i$ can be seen as the novel mid-level representation. To avoid the latent representation influencing by the non-uniform distribution of training images, we propose to enforce a Fisher discriminative regularization term [35] over the latent variables. The Fisher regularization can encode the label information of each image into the representation, which would further improve the discrimination of latent representation and the robustness to noises. Therefore, our objective function can be re-formulated as :

$$\underset{\{\mathbf{z}_i\},\Theta,\mathbf{W}}{\operatorname{argmin}} \sum_{i=1}^{N}(\alpha_i||\mathbf{z}_i - \Theta\varphi(\mathbf{X}_i)||_F^2) + ||\mathbf{y}_i - \mathbf{W}^T\mathbf{z}_i - \mathbf{b}||_2^2) \qquad (9)$$
$$+\beta||\Theta||_F^2 + \gamma||\mathbf{W}||_F^2 + \eta||\mathbf{z}_i||_1$$
$$+\lambda(tr(S_w(\mathbf{z}_i)) - tr(S_b(\mathbf{z}_i)) + ||\mathbf{z}_i||_F)$$
$$s.t.\ \mathbf{z}_i \geq 0,\ \forall i = 1, 2, ..., N,$$
$$S_w(\mathbf{z}_i) = \sum_{i=1}^{c}\sum_{\mathbf{z}_i\in\{\mathbf{z}_i\}_{i=1}^{n_c}}(\mathbf{z}_i - \mathbf{m}_c)(\mathbf{z}_i - \mathbf{m}_c)^T$$
$$S_b(\mathbf{z}_i) = \sum_{i=1}^{c} n_i(\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^T$$

where $\mathbf{m}_c \in \mathbb{R}^{M\times 1}$ is the mean representation of $\{\mathbf{z}_i\}_{i=1}^{n_c}$; n_c denotes the number of training images in the category c; $\mathbf{m} \in \mathbb{R}^{M\times 1}$ represents the mean vector of all the classes $\{\mathbf{z}_i\}_{i=1}^{N}$; and λ is trade-off parameter to determine the weighting of fisher regularization.

In the objective function, there exists λ, β, γ, and η hyper-parameters to be determined in the training phase. To determine these parameters, we conduct a validation set from each dataset in the experimental part. In fact, we find that the parameters do not generate much efforts on the experimental results. Thus, we select the value of each parameter when the objective function is convergence.

3.4 Optimization

The unknown variables in the objective includes $\{\mathbf{z}_i\}_{i=1}^{N}, \Theta, \mathbf{W}$, the minimization of which is not a jointly convex optimization problem. Therefore, we develop an alternating optimization algorithm to solve it by alternatively optimizing one variable at a time.

3.4.1 Update Θ. Discarding the constant terms, the objective function become a classic least squares minimization problem:

$$\underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{N}(\alpha_i||\mathbf{z}_i - \Theta\varphi(\mathbf{X}_i)||_F^2) + \beta||\Theta||_F^2. \qquad (10)$$

which can be efficiently computed. The closed form solution over the variables Θ is defined as:

$$\Theta = (\sum_i \alpha_i\mathbf{z}_i\varphi(\mathbf{X}_i)^T)(\beta\mathbf{I}_d + \sum_i \alpha_i\varphi(\mathbf{X}_i)\varphi(\mathbf{X}_i)^T)^{-1}, \qquad (11)$$

where $\mathbf{I}_d \in \{0, 1\}^{M\times M}$ is the identity matrix. The parameter Θ is initialized as an identity matrix.

Algorithm 1 Training procedure of our method

```
Input: $\{\varphi(\mathbf{X}_i)\}_{i=1}^{N}, \{\mathbf{y}_i\}_{i=1}^{N}, \lambda, \alpha_i, \beta, \gamma, \eta$
Initialize $\mathbf{\Theta} = \mathbf{I}_{M\times M}, \mathbf{z}_i = \varphi(\mathbf{X}_i)$
repeat
    Computing the latent attribute patterns discovery parameters $\mathbf{\Theta}$ according to Eq. 11 with $\mathbf{W}$, $\mathbf{b}$ and $\mathbf{z}_i$ fixed.
    Computing the classification parameter $\mathbf{W}$ and $\mathbf{b}$ according to Eq. 13 with $\mathbf{z}_i$ and $\mathbf{\Theta}$ fixed.
    Computing latent variables $\mathbf{z}_i$ according to Eq. 16 with $\mathbf{W}$ and $\mathbf{\Theta}$ fixed.
    Update the solutions $\mathbf{W}$, $\mathbf{b}$, $\mathbf{X}$ and $\mathbf{z}_i$.
until convergence or the maximize number of iterations
Output: Latent attribute patterns discovery parameter $\mathbf{\Theta}$, classification parameter $\mathbf{W}$ and $\mathbf{b}$.
```

3.4.2 Update $\mathbf{W}$ *and* $\mathbf{b}$. Given the latent variable $\{\mathbf{z}_i\}_{i=1}^{N}$, the configuration discovery parameter Θ, and the class labels of training samples $\{\mathbf{y}_i\}_{i=1}^{N}$, our objective function is formulated as:

$$\underset{\mathbf{W}}{\text{argmin}} \sum_{i=1}^{N} ||\mathbf{y}_i - \mathbf{W}^T\mathbf{z}_i - \mathbf{b}||_2^2 + \gamma||\mathbf{W}||_F^2, \tag{12}$$

This is also a convex optimization problem, and the closed form solution is:

$$\mathbf{W} = (\mathbf{Z}\mathbf{Z}^T + \gamma\mathbf{I}_d)^{-1}\mathbf{Z}\mathbf{Y}^T, \tag{13}$$

$$\mathbf{b} = \frac{1}{N}(\mathrm{Y} - \mathbf{W}^T\mathrm{Z})\mathbf{1}, \tag{14}$$

where $\mathrm{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N] \in \{0,1\}^{c\times N}$ and $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_N] \in \mathbb{R}^{M\times N}$. The size of $\mathbf{1}$ is $N \times 1$. Here we choose linear classification for computing efficiently, non-linear classifier can also be chosen *e.g.* SVM [5] with RBF kernel.

3.4.3 Update $\mathbf{Z}$. When the discovered parameter $\mathbf{\Theta}$ and prediction model parameter $\mathbf{W}$ and $\mathbf{b}$ are fixed, the optimization problem over the latent variables $\mathbf{z}_i$ could be divided into several subproblems. The latent advantage is that the solution could be computed in parallel for large scale computation:

$$\ell(\mathbf{z}_i) = \alpha_i||\mathbf{z}_i - \Theta\varphi(\mathbf{X}_i)||_F^2) + ||\mathbf{y}_i - \mathbf{W}^T\mathbf{z}_i - \mathbf{b}||_2^2 + \eta||\mathbf{z}_i||_1 \tag{15}$$
$$+\lambda(tr(S_w(\mathbf{z}_i)) - tr(S_b(\mathbf{z}_i)) + ||\mathbf{z}_i||_F).$$

To efficiently get the optimization solution for this objective function, we adopt a first-order projection gradient descent algorithm. The reason to choose this method is that the standard second-order quadratic solvers would be computationally expensive when the latent variables $\mathbf{z}$ is relatively large and the Hessian matrix grows quadratically. In each iteration, we compute the gradient of the objective function as:

$$\triangledown\ell(\mathbf{z}_i) = 2\mathbf{W}\mathbf{W}^T\mathbf{z}_i + 2\alpha_i\mathbf{z}_i - 2\alpha_i\mathbf{\Theta}\varphi(\mathbf{X}_i) - 2\mathbf{W}\mathbf{y}_i \tag{16}$$
$$+\eta + 2\mathbf{W}\mathbf{b} + \triangledown f(\mathbf{z}_i),$$

$$f(\mathbf{z}_i) = \lambda(tr(S_w(\mathbf{z}_i)) - tr(S_b(\mathbf{z}_i)) + ||\mathbf{z}||_F), \tag{17}$$

where $\triangledown f(\mathbf{z}_i)$ denotes the gradient of fisher discriminative term, which has been proven to be convex in [35]. For clarity, the whole optimization process of our proposed model has been summarized as shown in Alg. 1.

3.5 Testing

Given a new test image and the precomputed classifiers, we firstly compute the scores of attribute classifiers to develop the attribute description of the image $\mathbf{a}_{test}$. Then we extend such representation with embedding the co-occurrence of attributes under each category to get the co-occurrence embedding representation $\mathbf{X}_{test}$. Furthermore, the discriminative configuration of each class is embedded by $\mathbf{z}_{test} = \Theta\varphi(\mathbf{X}_{test})$ and the prediction score for each class is computed with $\mathbf{y}_{test} = \mathbf{W}^T\mathbf{z}_{test} + \mathbf{b}$. Finally, we get the label of test image $\mathbf{y}^\star$ by:

$$\mathbf{y}^\star = \underset{(r=1,2,\ldots,c)}{\text{argmax}} \ \mathbf{y}_{test}(r). \tag{18}$$

4 EXPERIMENT

4.1 Implementation details

To obtain the deep features by fast RCNN [15], we use the open-source Caffe [17]. We follow the fine-tune step to achieve the category models. In the step of fine-tune, we set the stochastic gradient descent with 0.9 momentum. We initiate learning rate to be 0.0001 and decrease it by 0.1 after finishing about 30 epochs. The weighting decay parameter is 0.0005. Our PC is configured with 2.8GHz CPU and GTX TITAN Black GPU. The mini-batch size of images is 256. The ImageNet ILSVRC-2012 dataset [8] is utilized to pre-train the CNN model by optimizing multinomial logistic regression objective function in the image classification task. This dataset contains about 1.2 million training images and 50,000 validation images. There is one main factor influencing the computational cost of our approach, which is the dimension of the graph representation $\mathbf{X}$ for each image. Its computational complexity for training is the minimum O(m) where m is the dimension of $\mathbf{X}$. As for the training time, for each iteration, it takes about 5 minutes using the unoptimized matlab&c++ code on an ordinary PC machine (Intel ES-2609 CPU and 64 GB RAM). Moreover, our proposed alternative iteration method converges in about 5 iterations with about 8% classification training error on the aPascal dataset.

4.2 Dataset

a-Pascal [9] consists of 20 classes with 6,340 training images, and 6,355 test images collected from PASCAL VOC 2008 challenge. On average each category has 317 images. Each image is assigned one of the 20 object class labels *e.g.* people, bird, cat and TV/monitor. Each image is also labeled by 64 binary attribute labels, *e.g.* "2D boxy", "has hair", "shiny". In addition, the attribute annotations is not labeled for the entire image. They are computed only for bounding box of the objects.

Unstructured Social Activity dataset [13] includes 8 different semantic class videos which are home videos of social occasions *e.g.* birthday party and music performance. This dataset contains around 100 videos for training and testing respectively. And, the visual and audio content of each video is manually annotated using 69 multi-model binary attributes, which can be summarized by five classes: actions, objects, scenes, sounds, and camera movement.

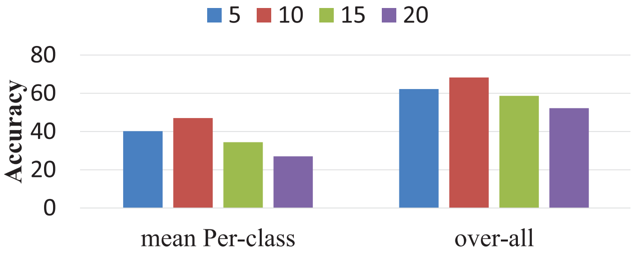
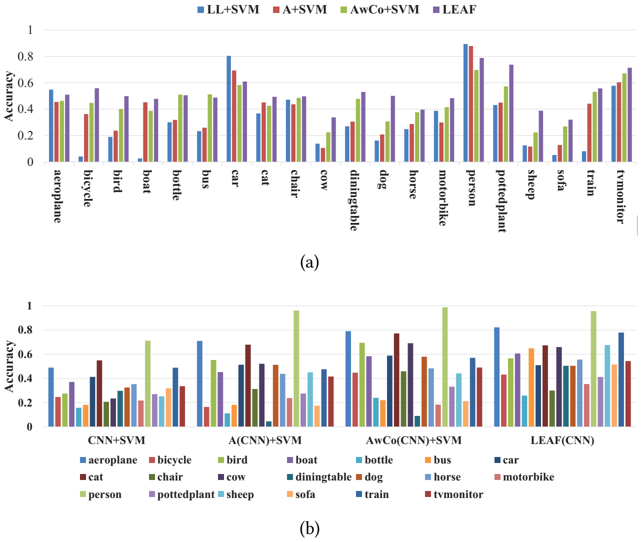


Figure 4: The performance of different cluster centers for visual classification. The number of clusters are $\{5, 10, 15, 20\}$



(a)

(b)

Figure 5: The comparison of APs for the 20 categories with three baseline methods on a-Pascal dataset. Best view in color and please zoom in for the clear comparisons.

4.3 Multi-attributes based image classification

On a-Pascal dataset, we use two kinds of low-level features to validate our proposed method. Each image is firstly represented as a $9,751$-dimensional feature vector extracted from information on color, texture, visual words, and edges, which are provided in [9]. Then, the 64 attribute classifiers are trained based on the raw-features of training images by the linear SVM [5]. The other kinds features are deep features which are extracted by Fast RCNN. Moreover, we use the fc7 full connected layer as the initialized feature representation, whose dimension is $4,096$ In this experiment setting, we use the same train/test splits provided by [9]. All the parameters of our method are determined by 3-fold cross-validation dataset. We report both overall and mean per-class accuracies following the existing work [9, 33, 34]. Two kind of evaluation metrics are selected as the criterion: Mean Per-class and overall accuracy. Mean Per-class precision is defined as the mean of average precision for all classes. While the overall accuracy is the number of all the right prediction of images. Furthermore, we also use average precision (AP) to evaluate the performance of each class.

Table 1: Object classification results on A-Pascal dataset

	Mean Per-class	Overall
Base feature + SVM [9]	35.5%	58.5%
Semantic attribute + SVM [9]	34.3%	56.1%
(Best results) [9]	37.7%	59.4%
Wang *et al.* [34]	50.84%	59.15%
Wang *et al.* [33]	44.82%	63.02%
LL+ SVM	31.78%	57.21%
A+ SVM	36.45%	53.80%
AwCo+ SVM	44.93 %	58.99%
LEAF	**52.13**%	**64.28**%
CNN+ SVM	33.51%	47.78%
A (CNN)+ SVM	41.02%	53.80%
AwCo (CNN)+ SVM	49.29%	65.29%
LEAF (CNN)	**56.40**%	**68.90**%

In this part, we firstly evaluate the number of cluster centers for each image. As discussed in Sec. 3.1, we cluster the object proposals to learn the attribute representation. Here we use the different number of centers to train the attribute classifiers, and then use these attribute scores to represent images. Based on different attribute representations, we evaluate the performance of category classifiers based on different centers as shown in Fig. 4.

From the results, we can observe that different number of centers $(5, 10, 15, 20)$ could influence the discriminative power of attribute representation. With the number of cluster center increasing, the classification results are firstly growing, but then decreasing. When the number of centers is 10, we achieve the best performance. The reasons can be summarized as that: Firstly, the small number of clusters would cause each center including some noise which can not easily remove by feature selection. While the large number of centers would introduce a set of noise into the training features.

We compare our method with three different baseline approaches to verify the advantages of our unified framework: our proposed method which used the Latent Extended Attribute Features as the image representation is denoted by **LEAF**. **LL** + SVM uses the low-level features as the image representation, and then the object classifiers are trained using this representation and the linear SVM. **A** + SVM denotes that the object classifiers are trained based on the attribute representation and SVM. **AwCo** + SVM indicates that we use the graph representation $\mathbf{X}_i$ as the image representation. Afterwards, we use the linear SVM to train the object models. In addition, several related work [9, 33, 34] are also introduced to compare with our proposed method. Moreover, we also conduct three baselines based on deep features. **CNN** + SVM is that we train the classifier of category based on FC7 layer. **A (CNN)** + SVM denotes that we train the attributes based on deep features. We also embed the co-occurrence into the attribute representation to develop a novel baseline named **AwCo (CNN)** + SVM. **LEAF (CNN)** represents that we explore the latent attribute patterns via deep features. The comparison results are summarized in Table 1.

From Table 1 we can observe that our proposed method **LEAF** achieves the best performance both on mean per-class (**52.13**%) and overall (**64.28**%) evaluations. Moreover, **LEAF** improves about 1.3% on the mean per-class accuracy and about **5.1**% on the overall accuracy comparing with [34]. While comparing with [33], our method

Table 2: The video classification results on Unstructured social activity dataset.

	KNN [12]	**SCA** [32]	**SLAS** [13]	**M2LATM** [12]	**LL** + SVM	**A** + SVM	**AwCo** + SVM	**LEAF**
I/10,A/69	26.8%	32%	40%	40.6%	30.69%	37.07%	42.41%	**47.48%**
I/10,A/R7	26.8%	25.6%	36%	38.3%	28.04%	29.12%	27.45%	**41.75%**

gains about 7.3% on the mean per-class accuracy and about 1.3% on the overall accuracy. We can find that **AwCo**+ SVM which embeds the co-occurrence obtains about 7.2% improvement on mean per-class compared with the best results in [9], while it achieves a little degeneration on the overall accuracy.

Benefiting from the discriminative ability of deep features, all the experimental results are better than these baselines based on low-level features. Comparing with [9], the performance of our proposed method (**CNN+SVM**) is degraded. While comparing with (**LL+SVM**), we can observe that it achieves improvement on Mean Per-class under the same SVM parameter settings. This may due to the scale problem and large overlapping between objects. We only have limited number of training images which can not cover all the situations. For these results based on attribute representation, we can observe that our proposed method achieve **56.4%** on mean per-class, and **68.9%** on overall performance. Comparing with **LEAF** with low-level features, the **LEAF(CNN)** gains about **4.7%** improvement on overall performance. We further show the average precision (AP) of each category comparing with our proposed baselines. As shown in Fig. 5, the results of **LL** + SVM and **A** + SVM are biased by the category "Person" and "Car", while **LEAF** achieves the relative uniform accuracies for all the categories. This can demonstrate that our proposed method is more robust to the imbalance distribution of training images comparing with these related methods. Wile the performance based on deep features, we can observe that most of the categories have significant improvement. This can demonstrate that the effectiveness of our method.

The encouraging experimental results can be explained by that: Firstly, the sequential steps to train the image classifier would be easily influenced by the unevenly distribution of training images on each categories, *e.g.* the "People" category in aPascal dataset, while our method is robust to the unbalanced training data. Secondly, we are not only focusing on training the discriminative classifiers, but also applying the extracted latent attribute structure for the test data. Thirdly, category-specific attribute classifiers can significant improve the discrimination of attribute as feature representation. Last but not the least, there exist latent structure correlation between attributes, which can help visual classification.

4.4 Video social activity classification

In this part, we further evaluate our method on attribute-based video classification problem. Three kinds of low-level features, which are SIFT, STIP, and MFCC, are extracted according to [18]. Then, the attribute classifiers are trained based on these low-level features by using one-vs-all methods. This dataset has the video-level attribute annotations. The train/test splitting is followed [18].

To validate the effectiveness of our proposed method, we train three baselines based on different feature representations, which are **LL** + SVM, **A** + SVM, and **AwCo** + SVM. The comparison results are summarized in Table 2. We also compare our method with other approaches on this dataset. **KNN** [12] uses the k nearest neighborhood as the classifiers. **SCA** [32] learns a generative model for class label and annotations based on topic model. **SLAS** [?] proposes to classify videos by introducing semi-latent attribute. **M2LATM** [12] classifies the videos based on the latent attributes.

Specifically, we randomly choose 10 videos ("I/10") of each class from the training data and all the 69 attributes ("A/69") are used to represent each video. Then, we reduce the number of attributes to describe the videos. 7 attributes are randomly selected from all the 69 attributes ("A/7") and the training images of each categories are still selected as "I/10". Table 2 summarizes the performance of our classification results and the comparison with the other seven methods. We can observe that our method (**47.48%**) significantly outperforms the baseline of directly using the graph representation (42.41%) on ("A/69, I/10"). Comparing with the state-of-the-art [12] on this dataset, our method gets about 7% and 4% improvements on these two tasks, respectively. Moreover, with the limited number of attributes and images ("A/R7, I/10"), our method can still achieve the better result comparing with related work. This further demonstrates the effectiveness of our unified framework.

5 CONCLUSION

In this paper, we have proposed a novel method to construct the extended attribute representation (named LEAF) via introducing the class-specific attribute graph representation. We firstly propose a multi-task learning method to select the common features across different categories to train the robust category-specific attribute classifiers. After that we introduce the attribute graph which are embedded the co-occurrence correlations between attributes. And then, based on the attribute graph, our goal is to discover the latent discriminative structures for visual classification. To achieve the discriminative representation, a novel framework is proposed which jointly learns the discriminative representation and the classifiers in a unified fashion. We have demonstrated that our proposed objective function could train the robust classifiers as well as discover the discriminative attribute patterns. Extensive experimental results on the challenge datasets have shown significant improvements for attribute based image classification, especially in case of limited training samples. In the future work, we will extend our work to mine the correlation between visual words. We also would like to reduce the complexity of our proposed method as the current representation grows exponentially with the number of attributes and categories.

6 ACKNOWLEDGEMENT

This work was supported by the National Key R&D Program of China (Grant No.2016YFC0801002), National Natural Science Foundation of China (No.61602464, 61422213, U1636214, U1605252, 6160 2337), Beijing Natural Science Foundation (No.4172068),Key Program of the Chinese Academy of Sciences (No. $QYZDB - SSW - JSC$003).

REFERENCES

[1] Shayan Modiri Assari, Amir Roshan Zamir, and Mubarak Shah. 2014. Video classification using semantic concept co-occurrences. In *CVPR*. 2529–2536.

[2] Tamara L Berg, Alexander C Berg, and Jonathan Shih. 2010. Automatic attribute discovery and characterization from noisy web data. In *ECCV*. 663–676.

[3] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. 2013. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *ACM Multimedia*. 459–460.

[4] Junjie Cai, Zheng-Jun Zha, Meng Wang, Shiliang Zhang, and Qi Tian. 2015. An Attribute-Assisted Reranking Model for Web Image Search. *TIP* 24, 1 (2015), 261–272.

[5] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2, 3 (2011), 27:1–27:27.

[6] Chao-Yeh Chen and Kristen Grauman. 2014. Inferring analogous attributes. In *CVPR*. 200–207.

[7] Huizhong Chen, Andrew Gallagher, and Bernd Girod. 2014. The Hidden Sides of Names-Face Modeling with First Name Attributes. *TPAMI* 36, 9 (2014), 1860–1873.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

[9] Alireza Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *CVPR*. 1778–1785.

[10] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. 2008. A discriminatively trained, multiscale, deformable part model. In *CVPR*. 1–8.

[11] Vittorio Ferrari and Andrew Zisserman. 2007. Learning visual attributes. In *NIPS*. 433–440.

[12] Y. Fu, T. Hospedales, T. Xiang, and S. Gong. 2014. Learning Multi-modal Latent Attributes. *TPAMI* 36 (2014), 303–316. Issue 2.

[13] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. 2012. Attribute learning for understanding unstructured social activity. In *ECCV*. 530–543.

[14] Yue Gao, Rongrong Ji, Wei Liu, Qionghai Dai, and Gang Hua. 2014. Weakly supervised visual dictionary learning by harnessing image attributes. *TIP* 23, 12 (2014), 5400–5411.

[15] Ross Girshick. 2015. Fast r-cnn. In *ICCV*. 1440–1448.

[16] Dinesh Jayaraman and Kristen Grauman. 2014. Zero-shot recognition with unreliable attributes. In *NIPS*. 3464–3472.

[17] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In *ACM Multimedia*. 675–678.

[18] Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C Loui. 2011. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ACM ICMR*. 29–36.

[19] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. 2009. Attribute and simile classifiers for face verification. In *ICCV*. 365–372.

[20] C. Lampert, Hannes Nickisch, and Stefan Harmeling. 2013. Attribute-based classification for zero-shot learning of object categories. *TPAMI* 36 (2013), 453–465. Issue 3.

[21] Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective.* MIT press.

[22] Devi Parikh and Kristen Grauman. 2011. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*. 1681–1688.

[23] Amar Parkash and Devi Parikh. 2012. Attributes for classifier feedback. In *ECCV*. 354–368.

[24] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. 2014. The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding. *IJCV* 108, 1–2 (2014), 59–81.

[25] Mohammad Rastegari, Abdou Diba, Devi Parikh, and Alireza Farhadi. 2013. Multi-attribute queries: To merge or not to merge?. In *CVPR*. 3310–3317.

[26] Jianbing Shen, Guo-Ping Liu, Jiann-Jong Chen, Yi Fang, Junfeng Xie, Yen-Ting Yu, and Shuo Yan. 2014. Unified Structured Learning for Simultaneous Human Pose Estimation and Garment Attribute Classification. *TIP* 23, 11 (2014), 4786–4798.

[27] Zhiyuan Shi, Yongxin Yang, Timothy M Hospedales, and Tao Xiang. 2014. Weakly supervised learning of objects, attributes and their associations. In *ECCV*. 472–487.

[28] Behjat Siddiquie, Rogerio S Feris, and Larry S Davis. 2011. Image ranking and retrieval based on multi-attribute queries. In *CVPR*. 801–808.

[29] Hyun Oh Song, Yong Jae Lee, Stefanie Jegelka, and Trevor Darrell. 2014. Weakly-supervised discovery of visual pattern configurations. In *NIPS*. 1637–1645.

[30] Yu Su and Frédéric Jurie. 2012. Improving image classification using semantic attributes. *IJCV* 100, 1 (2012), 59–77.

[31] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective search for object recognition. *IJCV* 104, 2 (2013), 154–171.

[32] Chong Wang, David Blei, and Fei-Fei Li. 2009. Simultaneous image classification and annotation. In *CVPR*. 1903–1910.

[33] Xiaoyang Wang and Qiang Ji. 2013. A unified probabilistic approach modeling relationships between attributes and objects. In *ICCV*. 2120–2127.

[34] Yang Wang and Greg Mori. 2010. A discriminative latent model of object classes and attributes. In *ECCV*. 155–168.

[35] Meng Yang, Lei Zhang, Xiangchu Feng, and David Zhang. 2014. Sparse Representation Based Fisher Discrimination Dictionary Learning for Image Classification. *IJCV* 109, 3 (2014), 209–232.

[36] Yang Yang, Zheng-Jun Zha, Yue Gao, Xiaofeng Zhu, and Tat-Seng Chua. 2014. Exploiting web images for semantic video indexing via robust sample-specific loss. *IEEE Transactions on Multimedia* 16, 6 (2014), 1677–1689.

[37] Xinge You, Ruxin Wang, and Dacheng Tao. 2014. Diverse Expected Gradient Active Learning for Relative Attributes. *TIP* 23, 7 (2014), 3203–3217.

[38] Felix X Yu, Rongrong Ji, Ming-Hen Tsai, Guangnan Ye, and Shih-Fu Chang. 2012. Weak attributes for large-scale image retrieval. In *CVPR*. 2949–2956.

[39] Junsong Yuan, Ming Yang, and Ying Wu. 2011. Mining discriminative co-occurrence patterns for visual recognition. In *CVPR*. 2777–2784.

[40] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, Yue Gao, and Tat-Seng Chua. 2013. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *ACM Multimedia*. 33–42.