Attribute-Guided Dictionary Learning

Wei Wang
DISI, University of Trento
wei.wang@unitn.it

Yan Yan
DISI, University of Trento
yan@disi.unitn.it

Nicu Sebe
DISI, University of Trento
sebe@disi.unitn.it

ABSTRACT

Attributes have shown great potential in visual recognition recently since they, as mid-level features, can be shared across different categories. However, existing attribute learning methods are prone to learning the correlated attributes which results in the difficulties of selecting attribute specific features. In this paper, we propose an attribute specific dictionary learning approach to address this issue. Category information is incorporated into our framework while learning the over-complete dictionary, which encourages the samples from the same category to have similar distributions over the dictionary bases. A novel scheme is developed to select the attribute specific dictionaries. The attribute specific dictionary consists of the bases which are only shared among the positive samples or the negative samples. The experiments on the Animals with Attributes (AwA) dataset show the effectiveness of our proposed method.

Categories and Subject Descriptors

H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation

Keywords

Attribute Learning, Dictionary Learning, Dictionary Bases

1. INTRODUCTION

In the real world, there exist numerous object categories. The performance of current machine learning approaches heavily relies on the sufficiency of training examples. However, the labeled data are often difficult and expensive to obtain. How to annotate images [29] and videos [28] is still an open problem. In order to leverage the knowledge of annotated categories to classify novel objects without training examples, visual attributes were proposed [5]. Visual attributes are the middle-level descriptors which bridge the low-level features and high-level concepts such as object categories.

The most important property of the attributes, such as color and shape, is that they can be transferred among different object categories. Zero-shot learning [13] is proposed as an attribute-based classification method based on this property. First, attribute classifiers are pre-learned independently from related objects. Then the target object can be recognized based on its attribute representation, which requires no training examples.

Jayaraman *et al.* [11] pointed out that the performance of attribute classifiers could be improved through feature selection because of the intrinsic relations between attributes and features. For example, color attributes (*red, green, yellow, etc.*) can be better trained on the dimensions corresponding to color histogram bins, wheres texture attributes (*furry, silky, etc.*) prefer texture features. Current attribute learning methods usually map the low-level features directly to the attributes. However, the dimension of low-level feature vectors is usually very high because of the concatenation of various features, such as SIFT, Color SIFT and HOG.

Conventional methods perform feature selection by incorporating both l_1-norm & l_2-norm or $l_{2,1}$-norm regularizer, which encourage sparsity selection of features and incorporate the attributes' correlation at the same time [8, 9, 11]. For instance, l_1-norm encourages feature competition among groups, l_2-norm encourages feature sharing among groups, and $l_{2,1}$-norm encourages intra-group feature sharing and inter-group competition. Regardless of regularizer types, the underlying intuition remains the same, i.e., encourage the semantically close attributes to share similar feature dimensions. The semantic correlation is either measured according to the semantic distance mined from the web, e.g., using WordNet [7], or from attributes' co-occurrence probability as proposed by Han *et al.* [9]. However, it is usually hard to judge to what extent the semantic closeness can reflect the visual appearance similarity, and there is no guarantee that the semantically close attributes are visually similar. For example, the similarity between *cat* and *man* is 0.14 and 0.11 between *cat* and *dolphin*, based on the Leacock-Chodorow similarity measurement from WordNet [7]. However, we could not say that the appearance of *man* is more similar to *cat* compared with *dolphin*. Moreover, feature selection over the low-level features discards the structure information as each feature dimension is treated independently.

ICMR'15, June 23–26, 2015, Shanghai, China.

http://dx.doi.org/10.1145/2671188.2749337.

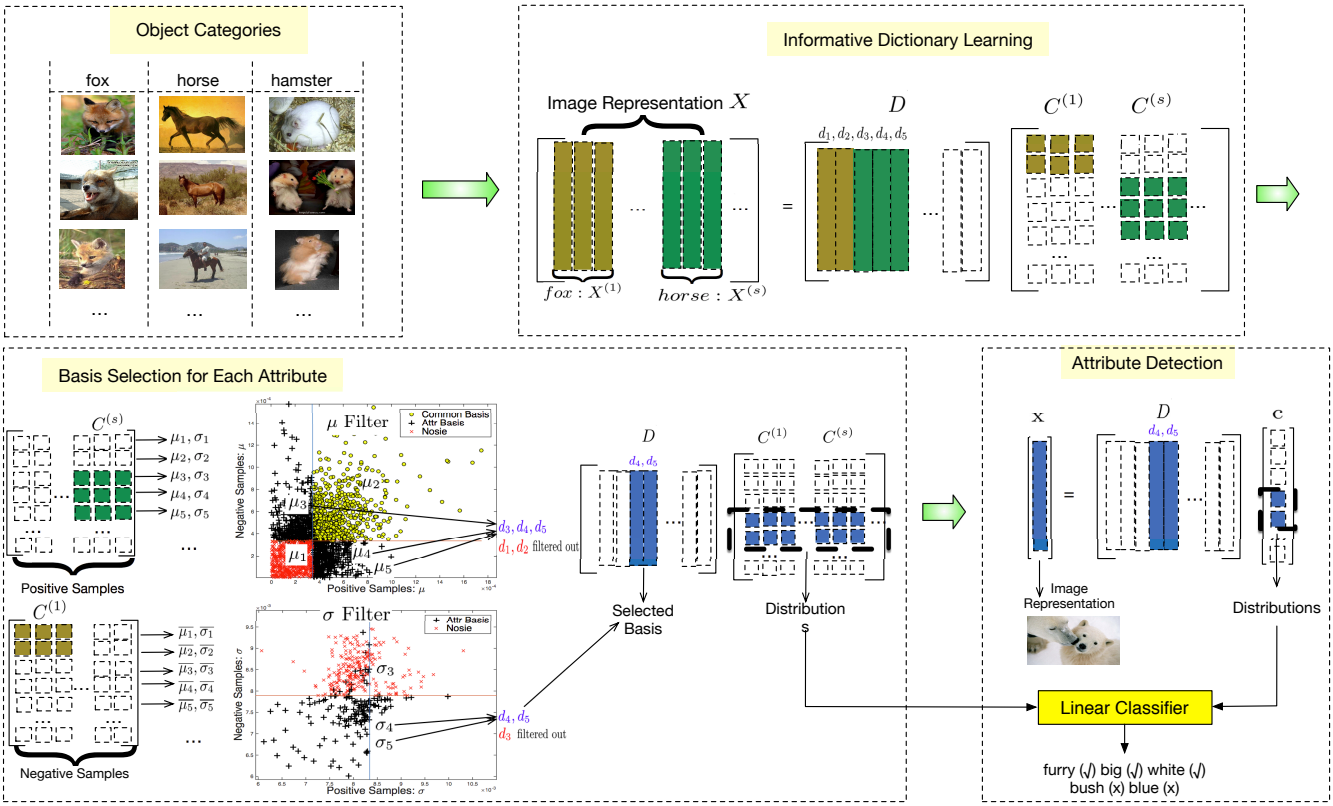

Figure 1: Overview of Proposed Framework. (top) An informative dictionary is learned by incorporating category labels. A novel multi-vector FISTA algorithm is designed to learn the new dictionary. (bottom left) two filters: μ-Filter & σ-Filter are used to select a set of robust and representative attribute specific bases to reconstruct each attribute. (bottom right) Final attributes are predicted by linear SVM classifiers using their distributions over the attribute specific bases.

To address this issue, in this paper, we propose a novel attribute specific dictionary learning approach. Fig.1 shows the overview of our proposed framework. Unlike the conventional methods which map the low-level features directly to the attribute, we adopt feature selection over dictionary, as a dictionary is expected to capture higher-level structure of images [15]. First, an informative dictionary is constructed by incorporating the label information of training data. Second, we perform basis selection according to distribution statistics for each attribute. Attribute specific bases are selected if only the positive or only the negative examples have large and stable distribution over the bases. The larger the distribution is, the more representative the bases will be. Another important factor which needs to be considered is that the robustness of a basis can be quantified by standard deviation. The smaller the standard deviation is, the more robust the basis will be. Therefore, in our proposed method, two filters are designed to select the attribute specific bases, namely, μ-Filter and σ-Filter. The μ-Filter will select the representative bases first. Then the σ-Filter will measure the robustness of the bases and judge whether they are qualified to be selected as attribute specific bases. Common bases are marked if both the positive and negative examples have large distribution over them. The common bases are only used for the reconstruction while the attribute specific bases are used both for reconstruction and attribute classifier training. Finally, the attributes are predicted by a linear SVM classifier using the distribution over the attribute specific bases. To sum up, this paper makes the following contributions:

- A novel informative dictionary learning method is proposed by incorporating label information from the data. The training data with the same labels are encouraged to have similar distributions. A novel multi-vector FISTA algorithm is designed to solve the problem.
- A new dictionary basis selection method is proposed to perform feature selection. Two filters, namely μ-Filter and σ-Filter are designed to select the robust and representative bases for each attribute.

This paper is organized as follows. Section 2 reviews related work. Section 3 introduces our proposed method. Experiments are described in Section 4, while Section 5 concludes this paper.

2. RELATED WORK

In this section, we review the related work on attribute learning, feature selection for attributes and dictionary learning for attributes.

2.1 Attribute Learning

Attributes usually provide more details about an image. In some situations, people may want to know not only the

object categories (*e.g., cat, dog, bike*), but also the attributes (*e.g., is silky, has legs, is cute*) in an image. As a type of mid-level features, attributes have shown great potential in object recognition tasks [4, 5]. Lampert *et al.* [13] proposed zero-shot learning to predict unseen objects based on attribute learning. Farhadi *et al.* [6] proposed describing an image based on the generated structured semantic triples <*object, action, scene*>. Han *et al.* [10] proposed a hierarchical tree-structured semantic unit to describe an image at different semantic levels (*attribute level, category level, etc*). Since attribute learning is widely used, the performance of an attribute classifier is essential to its success.

2.2 Feature Selection for Attributes

The conventional methods learn the attribute classifier by mapping the low-level image features directly to semantic attributes, in which regularization plays an important role. The most important function of the regularizer is to perform feature selection in order to prevent over-fitting.

There exist many different attribute groups, such as person-related attributes (*e.g., is male, has hat, has glasses*), scene attributes (*e.g., trees, clouds, leaves*) and animal attributes. In animal attributes group, there are also subgroups, such as textures (*e.g., stripes, furry, spots*), part-of-body (*horns, claws, tusks*) and colors (*black, white, blue*). Jayaraman *et al.* [11] pointed out that the attribute classifiers would have different performances when different features were used because of the intrinsic relations between attributes and features. For example, color attributes can be better trained on the dimensions corresponding to color histogram bins, whereas texture attributes prefer texture features.

Thus, feature selection is an important process to improve attribute classifiers' performance. Many works imposed feature selection directly on the low-level features by using regularizers, such as l_1-norm combined with l_2-norm, $l_{2,1}$-norm [11, 19], or $l_{2,p}$-norm [26, 25] as well as different loss functions, such as *linear regression* or *logistic regression*. Most current works showed that attribute classifiers' performance can be further improved by incorporating correlation between attributes into the regularizer. To capture the correlation between attributes, various methods are proposed. Han *et al.* [9] constructed a connected graph to represent the correlation between each pair of attributes and incorporated the relation into l_1-norm regularizer. Chang *et al.* [3, 2] proposed mining the correlation between semantic labels from the shared common structures of the training data in a low-dimensional subspace. Song *et al.* [17, 18] proposed hashing methods for large-scale image and video retrieval by imposing semantic attribute information. Jayaraman *et al.* [11] proposed using $l_{2,1}$-norm regularizer which encouraged intra-group feature sharing by l_2-norm and discouraged inter-group feature sharing by l_1-norm. Regardless of the regularizer types, all these methods depend on the regularizer for doing feature selection.

2.3 Dictionary Learning for Attributes

Dictionary learning has been successfully applied into supervised learning tasks. This approach significantly improves classification performance as shown in Raina *et al.* [15]. More recently, dictionary learning has been applied into event detection problems [27]. It can also be used for image clustering tasks. Ramirez *et al.* [16] proposed building a dictionary for each individual cluster. The new data are assigned to the cluster whose corresponding dictionary can reconstruct the data with the minimum reconstruction error. However, the dictionaries of different clusters have intersections and their associated reconstruction coefficients are usually large. The discriminative power can be enhanced by neglecting the coefficients of the shared bases.

Some work tried to bridge attribute and dictionary learning. Feng *et al.* [8] proposed an adaptive dictionary learning method for object recognition. Each image is reconstructed by a linear binary combination of dictionary bases, and each basis is regarded as an attribute. Attributes can also be also applied into action recognition. Instead of employing 3D models [20] or body segmentation method [21], Qiu *et al.* [14] proposed learning a compact dictionary for action recognition, in which every basis is regarded as an action attribute. However, all these attributes have no semantic meanings, and thus can hardly be generalized to a novel action. Besides, the dictionary is usually trained by using unlabeled data [16, 15]. When labeled data are available, a more informative dictionary can be learned, which is expected to encourage the data with the same label to have similar distributions. In our work, this is implemented by a special regularizer and a novel multi-vector FISTA is adopted to solve the problem.

3. ATTRIBUTE GUIDED DICTIONARY LEARNING

Inspired by [14], we propose selecting attribute specific bases for each attribute. First, an informative dictionary is learned by incorporating label information of training data. Then the attribute specific bases are selected. It is worth noticing that two types of attribute specific basis are considered: the basis that is only shared among the positive examples, and the one that is only shared among the negative examples. These two basis types are named as *positive stimulus basis* which reflects what the attribute *has* and *negative stimulus basis* which reflects what the attribute does *not have*.

3.1 Informative Dictionary Learning

The classical dictionary learning model only considers reconstruction error and sparsity of distribution. The model is defined as follows:

$$\min_{\mathbf{D},\mathbf{C}} \|\mathbf{X} - \mathbf{D}\mathbf{C}\|_F^2 + \lambda \sum_{i=1}^{N} \|\mathbf{c_i}\|_1$$

where $\mathbf{X} \in \mathbb{R}^{M \times N}$, M is the dimension of training data, N is the number of data, $\mathbf{D} \in \mathbb{R}^{M \times L}$ is the dictionary, L is the number of bases, $\mathbf{C} \in \mathbb{R}^{L \times N}$ is the distribution matrix of training data, and $\mathbf{c_i}$ is the i-th column of $\mathbf{C}$, l_1-norm is the lasso constraint which encourages sparsity, and λ balances the trade-off between the reconstruction error and the sparsity.

Ramirez *et al.* [16] proposed learning multiple dictionaries for multiple categories to better embed the class information. Instead of learning multiple dictionaries, we only learn one informative dictionary for all categories which encourages the data with same labels to have similar distributions. We propose solving the following novel optimization problem:

$$\min_{\mathbf{D},\mathbf{C}} \|\mathbf{X} - \mathbf{D}\mathbf{C}\|_F^2 + \alpha \sum_{i=1}^{N} \|\mathbf{c_i}\|_1 + \beta \sum_{s=1}^{K} \|\mathbf{C^{(s)}} - \overline{\mathbf{C^{(s)}}}\mathbf{E_s}\|_F^2 \quad (1)$$

The third term $\sum_{s=1}^{K} \|\mathbf{C^{(s)}} - \overline{\mathbf{C^{(s)}}}\mathbf{E_s}\|_F^2$ helps minimizing the distribution variances within each category. K is the number of categories. $\mathbf{C^{(s)}} = [\mathbf{c_1^{(s)}}, \mathbf{c_2^{(s)}}, ..., \mathbf{c_k^{(s)}}]$ denotes the distribution of data from category s. $\overline{\mathbf{C^{(s)}}}$ is the mean of $\mathbf{C^{(s)}}$. $\mathbf{E_s} = [1, 1, ...]_{1 \times s_k}$, where s_k is the number of data from category s. α balances the reconstruction error against sparsity while β provides the trade-off between the measurements and variance penalty. By encouraging the data from the same category to have similar sparse representations, the model will focus on learning the shared attributes among the instances in the same category. In addition, learning all the bases in the same dictionary, instead of multiple dictionaries, allows bases sharing across different categories. Thus, the attribute bases can be isolated by searching the shared bases across different categories whom containing the same specific attribute.

3.1.1 Optimization

To solve the proposed optimization problem, we decompose the objective into sub-objectives and then optimize the sub-objectives alternatively. We adopt a novel multi-vector FISTA algorithm to solve the sub-objectives. The details are shown in Algorithm 1.

Algorithm 1 Solution Structure

Initialization: $\mathbf{D} \leftarrow \mathbf{D_0}$, $\mathbf{C} \leftarrow \mathbf{C_0}$
repeat
  fix D, update C:
  for $C^{(s)} \in C$ **do**
    $ratio \leftarrow 1$
    while $ratio > threshold$ **do**
      run ***Multi-Vector FISTA***
      update $ratio$
    end while
  end for
  fix C, update D
until converges

(1) Fix C, Optimize D: By setting the derivative of Eqn.(1) with respect to $\mathbf{D}$ equal to $\mathbf{0}$, we obtain,

$$(\mathbf{DC} - \mathbf{X})\mathbf{C^T} = \mathbf{0} \Rightarrow \mathbf{D} = \mathbf{XC^T}(\mathbf{CC^T})^{-1}$$

In case $\mathbf{CC^T}$ is singular, we use the following equation to update $\mathbf{D}$.

$$\mathbf{D} = \mathbf{XC^T}(\mathbf{CC^T} + \lambda\mathbf{I})^{-1}$$

λ is a small number and it guarantees that the matrix $\mathbf{CC^T} + \lambda\mathbf{I}$ is invertible when $\mathbf{CC^T}$ is singular.

(2) Fix D, Optimize C: To update $\mathbf{C}$, we decompose the objective into a group of sub-objectives. Each sub-objective corresponds to one category. The following are important steps of decomposition. The first term in Eqn.(1) can be re-organized as:

$$\begin{aligned} \|\mathbf{X} - \mathbf{DC}\|^2 &= \sum_{i=1}^{N} \|\mathbf{Dc_i} - \mathbf{x_i}\|^2 \\ &= \sum_{s=1}^{K} \|\mathbf{DC^{(s)}} - \mathbf{X^{(s)}}\|^2 \end{aligned} \tag{2}$$

For the second term, we have:

$$\sum_{i=1}^{N} \|\mathbf{c_i}\|_1 = \sum_{s=1}^{K} \|\mathbf{C^{(s)}}\|_1 \tag{3}$$

By putting Eqn.(2) and Eqn.(3) back into the objective function of Eqn.(1), we can obtain the new form of the objective function:

$$\min_{\mathbf{D},\mathbf{C}} \sum_{s=1}^{K} L(\mathbf{D}; \mathbf{C^{(s)}}; \mathbf{X^{(s)}})$$

where

$$L(\mathbf{D}; \mathbf{C^{(s)}}; \mathbf{X^{(s)}}) = \|\mathbf{DC^{(s)}} - \mathbf{X^{(s)}}\|_F^2 + \alpha\|\mathbf{C^{(s)}}\|_1 + \beta\|\mathbf{C_{(s)}} - \overline{\mathbf{C^{(s)}}}\mathbf{E_s})\|^2$$

Note that when $\mathbf{D}$ is fixed, $L(\mathbf{D}; \mathbf{C^{(s)}}; \mathbf{X^{(S)}})$ is independent from each other with respect to s. Then the objective can be written as:

$$\min_{\mathbf{C}} \sum_{s=1}^{K} L(\mathbf{D}; \mathbf{C^{(s)}}; \mathbf{X^{(s)}}) = \sum_{s=1}^{K} \min_{\mathbf{C^{(s)}}} L(\mathbf{D}; \mathbf{C^{(s)}}; \mathbf{X^{(s)}})$$

Thus the original objective function can be decomposed into a group of sub-objective functions with respect to each category. Beck *et al.* [1] proposed using the Fast Iterative Soft-Thresholding Algorithm (FISTA) algorithm to solve the classical dictionary learning problem, which is proved to converge very fast. As introduced by Beck *et al.* [1], a soft-threshold step is incorporated into FISTA to guarantee the sparseness of the solution. FISTA converges in function values as $O(1/k^2)$, in which k denotes the iteration times, while for the traditional ISTA method, the complexity is $O(1/k)$.

However, the original FISTA can not be applied to our model because of the third term in Eqn.(1). The third term makes the $\mathbf{c_i}$ and $\mathbf{c_j}$ within the same category become dependent on each other. Thus, $\mathbf{c_i}$ and $\mathbf{c_j}$ must be updated simultaneously in order to make the whole system converge. We propose a novel multi-vector FISTA algorithm to tackle the problem.

3.1.2 Multiple-Vector FISTA

Our objective function is quite different from the classical dictionary learning objective. In the classical model, the distributions of different data are independent from each other. Thus when the overall objective is decomposed into sub-objectives, each sub-objective can be optimized independently. However, in our model, we require the training data $\mathbf{x}$ which belong to the same group $\mathbf{X^{(s)}}$ to have similar distributions. Thus, the new sub-objectives are not independent. The sub-objective with respect to each training data is as follows,

$$\|\mathbf{Dc} - \mathbf{x}\|^2 + \alpha\|\mathbf{c}\|_1$$

The new sub-objective in our model is as follows,

$$\sum_{c \in C^{(s)}} \|\mathbf{Dc} - \mathbf{x}\|^2 + \alpha\|\mathbf{c}\|_1 + \beta\|\mathbf{c} - \frac{1}{N}\sum_{c_k \in C^{(s)}} \mathbf{c_k}\|^2$$

From the equation above, we can find that, for training data $\mathbf{x} \in \mathbf{X^s}$, its distribution $\mathbf{c}$ ($\mathbf{c} \in \mathbf{C^s}$) depends on other $\mathbf{c_k}$ ($\mathbf{c_k} \in \mathbf{C^s}$). Thus, the sub-objectives cannot be optimized independently. Based on FISTA, we propose a multi-vector FISTA to optimize the sub-objectives from the same group simultaneously. The sub-objectives are grouped together if the training data belong to the same group.

The gradient of $\mathbf{c}$, $\frac{\partial \mathbf{F}}{\partial \mathbf{c}} = 2\mathbf{D}^{\mathbf{T}}(\mathbf{Dc} - \mathbf{x})$, is replaced by

$$\frac{\partial \mathbf{F}}{\partial \mathbf{C^{(s)}}} = \begin{cases} ... \\ \frac{\partial \mathbf{F}}{\partial \mathbf{c_j}} = 2\mathbf{D}^{\mathbf{T}}(\mathbf{Dc_j} - \mathbf{x}) + 2\beta(\mathbf{c_j} - \frac{1}{N}\sum_{c_k \in C^{(s)}} \mathbf{c_k}) \\ ... \end{cases} \tag{4}$$

Then when updating $\mathbf{C^{(s)}}$, all $\mathbf{c_j} \in \mathbf{C^{(s)}}$ are updated simultaneously for $j = 1, ..., i_k$.

$$\mathbf{c_j} := \mathbf{c_j} - \gamma \frac{\partial \mathbf{F}}{\partial \mathbf{c_j}}$$

This updating procedure of $\mathbf{C^{(s)}}$ continues until it converges. To judge whether all the $\mathbf{c_j}$ in the same category converge or not, we refer to the metric *ratio*, which is defined as:

$$ratio = \min_{\mathbf{c_j} \in \mathbf{C^{(s)}}} \|\mathbf{c_j} - \widehat{\mathbf{c}}_\mathbf{j}\|^2 / \|\widehat{\mathbf{c}}_\mathbf{j}\|^2$$

in which $\widehat{\mathbf{c}}_\mathbf{j}$ denotes the updated value of $\mathbf{c_j}$.

If $ratio < threshold$, the update procedure for the category will be terminated. For each of the categories, we run the same procedure. In Algorithm 1, line 4 to line 10 represent the pseudo-code to update $\mathbf{C}$.

3.2 Basis Selection

To judge whether a basis belongs to **common bases** or to **attribute specific bases**, we rely on the statistics of distribution $\mathbf{C}$. *Common Bases* are the bases on which both the positive and negative examples have large distributions. *Attribute Specific Bases* are the bases on which only the positive or only the negative examples have large and stable distributions. Two metrics, the mean μ and the standard deviation σ, are used to describe the distribution of samples over the bases. Thus, we design two filters for basis selection: μ-Filter and σ-Filter.

For attribute specific bases, two types of attribute specific bases are considered, the positive stimulus bases, which are only shared among the positive examples, and negative stimulus bases, which are only shared among the negative examples. *Positive stimulus basis* represents what the attribute *has* while *negative stimulus basis* represents what the attribute does *not have*.

Let $c_{i,j}$ denotes the j-th sample's value over i-th basis. The mean of positive samples over the i-th basis is $\mu_i = \frac{1}{|P|}\sum_{j \in P} c_{i,j}$. P is the set of positive samples and $|\ P\ |$ is the cardinality of the set. Similarly, we have $\overline{\mu_i} = \frac{1}{|N|}\sum_{j \in N} c_{i,j}$ for the negative set N. Then the criteria for basis selection is shown in Fig.2. μ_P, $\overline{\mu_N}$, σ_P and $\overline{\sigma_N}$ are threshold values that control the amount of selected bases for each type.

μ-Filter The candidates of positive stimulus bases are the ones which are located in **region 2** in the μ-Filter section of Fig.2. The candidates of negative stimulus are located in **region 4** in the μ-Filter. The common bases are located in **region 1**. Among those candidates, only the robust candidates will finally be selected as attribute specific bases.

σ-Filter σ-Filter is the second layer filter. Given a candidate of positive stimulus basis, it will be selected as a positive stimulus if the standard deviation of positive examples over the basis is small, as shown in **region 4** & **region 3** in the σ-Filter section of Fig.2. The robust negative stimulus bases are selected in the same way, located in **region 2** & **region 3**.

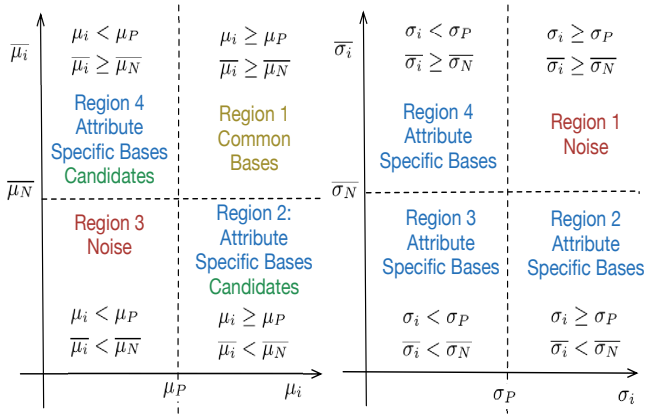

Figure 2: μ-Filter & σ-Filter

3.3 Attribute Classifier

After the attribute specific basis selection, we adopt linear SVM as attribute classifiers. The distributions over the attribute specific bases are selected as the training data for the classifiers. To detect whether an image contains a certain attribute, the image is first reconstructed by all the bases. Then its distribution over the attribute specific bases will be used to perform attribute detection.

To evaluate the performance of our method, we use F_1 score, which is the harmonic mean of precision and recall:

$$F_1 = 2\frac{precision \cdot recall}{precision + recall}$$

After obtaining the F_1 scores of multiple attributes, their average value is used as the evaluation metric.

4. EXPERIMENTS

In this section, we conduct extensive experiments to evaluate our proposed method.

4.1 Dataset

'Animal with Attributes' (AwA) dataset [12] is used in our experiment. This dataset contains 50 animal categories, which are separated into 2 parts: 40 seen animal categories and 10 unseen animal categories. The attribute classifiers are trained by the data instances from the seen categories and are tested on the data instances from the unseen categories. 85 semantic attributes are defined in the dataset, which are grouped into 9 groups (*color, texture, shape, etc.*). The attributes are mapped to the categories according to the attribute-category matrix. Fig.3 shows some examples from the AwA dataset. We evaluate our method with different types of features.

4.2 Parameter Settings

In the category dictionary learning phase, α and β are tuned from $[10^{-3}, 10^{-2}, ..., 10^3]$. Dictionary size is tuned from $[1, 1.5, ..., 3] \times 10^3$. In the basis selection phase, the threshold values μ_P, $\overline{\mu_N}$, σ_P, $\overline{\sigma_N}$ control the ratio of bases that are selected for each attribute. The ratio is tuned

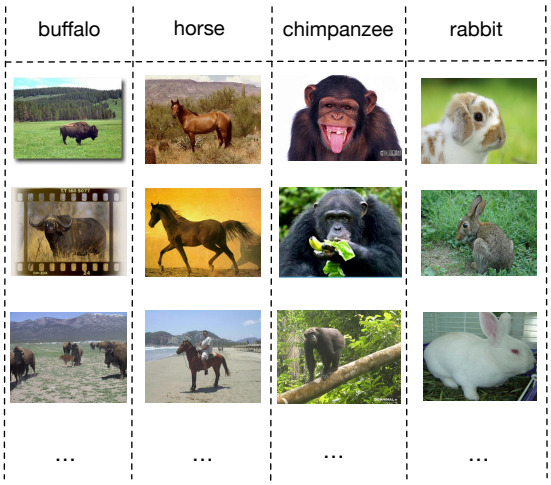

Figure 3: Examples from the AwA dataset

from $[10\%, 20\%, ..., 100\%]$. In the attribute classifier training phase, the penalty parameter C in the SVM classifier is tuned from $[10^{-3}, 10^{-2}, ..., 10^{3}]$.

4.3 Results

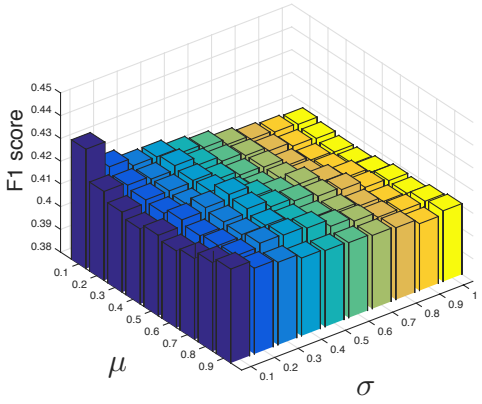

Figure 4: F_1 Score over μ & σ

Fig.4 shows the grid plot of F_1 score with respect to different filter parameters when the dictionary size is set to be 2000. μ-Filter and σ-Filter control the amount of selected bases. The ratio of the selected bases is coarsely grained, ranging from 10% to 100%. The performance is measured on the unseen object categories. The value of the bar is the mean F_1 score of all the 85 attributes. From Fig.4, we observe that when more bases are selected either by the μ-Filter or by the σ-Filter, the mean F_1 score tends to decrease, even though it does not decrease monotonically. The maximum F_1 score is obtained when both the μ-Filter and the σ-Filter only select 10% bases. From this observation, we can conclude that, the basis selection improves the performance of attribute detectors, and the best ratio of the basis selection lies between 0% to 20% which could be mined out by doing a fine-grained search of the ratio.

Fig.5 shows the scatter plot of basis selection with the μ-Filter. The two decision boundaries control the ratio of the selected candidates of the attribute specific bases. Then

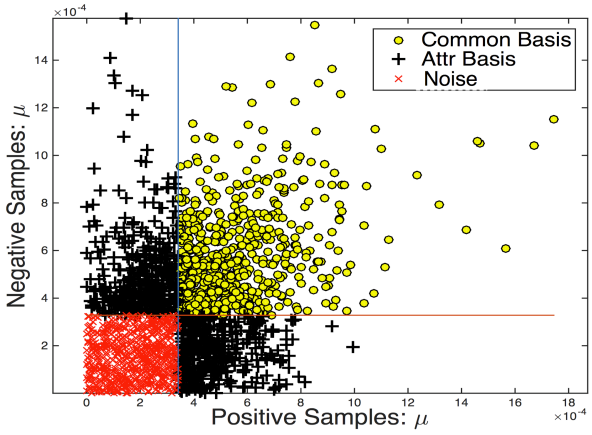

Figure 5: Scatter Plot of bases in μ-Filter

the selected bases are filtered again with the σ-Filter, as shown in Fig.6. Thus, the robust and representative bases are obtained. The decision boundaries in the μ-Filter are determined by two threshold values, μ_P, $\overline{\mu_N}$ which correspond to the mean of the distribution of the positive samples, and to the mean of the distribution of the negative samples respectively. The two boundaries divide the bases into four regions. However, only the bases in the upper-left region and the lower-right region are selected as representative bases. The bases in the upper-left region represent what the attribute does *not have*. The bases in the lower-right region represent what the attribute *has*. The bases in lower-left region are regarded as noises as both positive and negative samples have small distributions over them. By setting the boundaries to different values, different amount of bases can be selected.

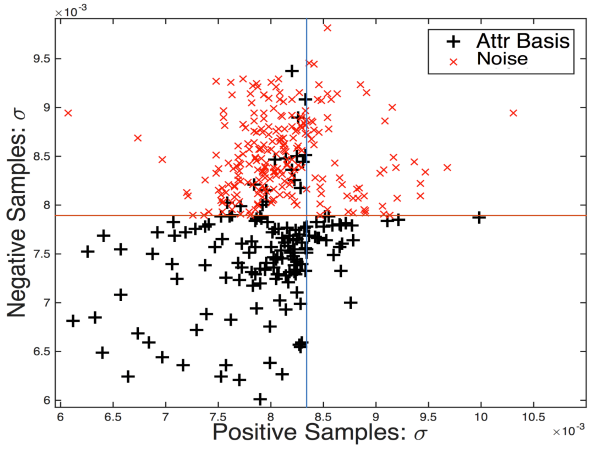

Figure 6: Scatter Plot of bases in σ-Filter

Fig.6 illustrates the second layer filter, namely, the σ-Filter. The σ-Filter is applied to judge whether the representative bases are robust and thus can be selected as attribute specific bases. The decision boundaries in the σ-Filter are determined by two threshold values, σ and $\overline{\sigma_N}$ which correspond to the standard deviation of positive samples, and the

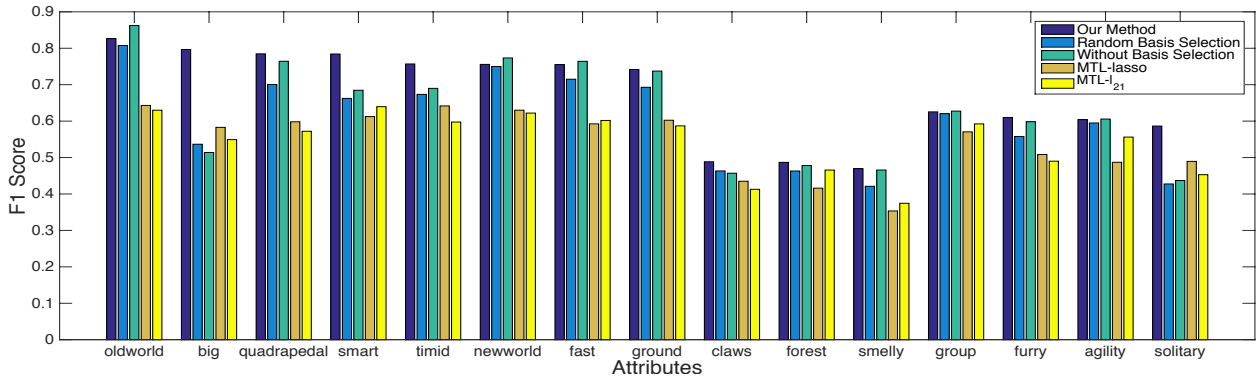

Figure 7: F_1 Score of Attribute Detection on Unseen Animal Categories

Table 1: Performance Comparison with Baselines Based on Different Features

Methods	Features							
	SIFT		Color-SIFT		Pyramid-HOG		DeCaf	
	F_1 score	Precision	F_1 score	Precision	F_1 score	Precision	F_1 score	Precision
Without Basis Selection	0.4011	0.4165	0.3981	0.4064	0.4270	0.4193	0.4163	0.4346
Random Basis Selection	0.4122	0.4140	0.4070	0.4071	0.4273	0.4233	0.4159	0.4179
MTL *lasso*	0.4088	0.4077	0.4401	0.4187	0.4311	0.3843	0.4177	0.3797
MTL $l_{2,1}$-norm	0.4107	0.4097	0.4167	0.4112	0.4166	0.3860	0.4432	0.4386
Our method	**0.4789**	**0.4493**	**0.4465**	**0.4348**	**0.4815**	**0.4328**	**0.4752**	**0.4568**

standard deviation of negative samples respectively. If either the positive samples or the negative samples have large mean and small standard deviations over the representative bases, the bases will be selected as attribute specific bases. Otherwise, they will be filtered out.

After performing basis selection for the 85 attributes, the next step is to train the attribute classifiers and apply these classifiers to unseen categories. Three baselines are used to compare with our method. The first is the one without basis selection. The second is the one which performs feature selection randomly. We also use multi-task learning (MTL) methods [24, 22, 23] with lasso regularizer and $l_{2,1}$-norm regularizer as baselines.

Fig.7 shows the F_1 score for some attributes using SIFT feature. It shows that our method outperforms the other methods in 64 out of 85 attributes. Table 1 shows the comparison with other baselines on attribute detection using different features. We measure the performance of an attribute detector by averaging the F_1 score of all the 85 attributes, as well as the precision. We can see from Table 1 that the performance of our method is better than the others.

We also study the convergence of our algorithm. Fig.8(a) shows the convergence curve of the overall function. Fig.8 (b) is the log plot of the loss when updating $C^{(s)}$ for five categories. It shows that all the five sub-objectives converge very fast.

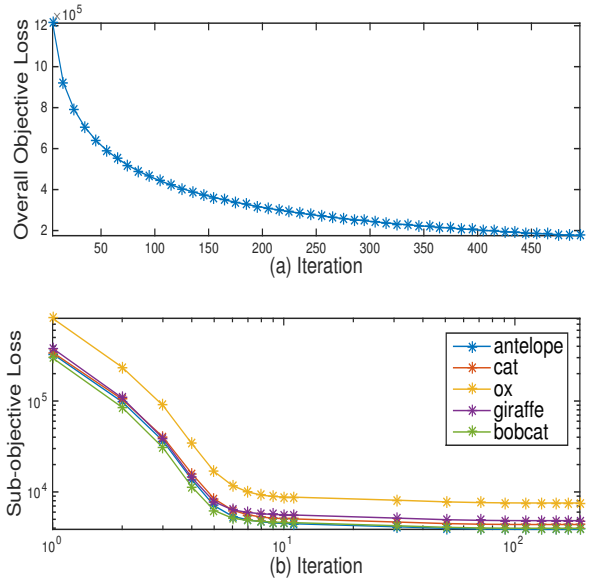

Figure 8: Convergence curve of the overall function

5. CONCLUSIONS

In this paper, we propose a novel attribute guided dictionary learning method to improve the performance of attribute detectors. First, an informative dictionary is learned utilizing the label information. Then we design a multi-vector FISTA algorithm to solve the problem. Specifically, we propose μ-Filter and σ-Filter to select a group of robust and representative attribute specific bases for each attribute. The extensive experimental results show that our proposed method outperforms other important baselines.

Acknowledgment: This research is partially supported by the xLiMe EU project.

6. REFERENCES

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[2] X. Chang, F. Nie, Y. Yang, and H. Huang. A convex formulation for semi-supervised multi-label feature selection. In *AAAI*, 2014.

[3] X. Chang, H. Shen, S. Wang, J. Liu, and X. Li. Semi-supervised feature analysis for multimedia annotation by mining label correlation. In *PAKDD*. 2014.

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

[5] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. CVPR, 2009.

[6] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. ECCV, 2010.

[7] I. Feinerer and K. Hornik. *WordNet: WordNet Interface*, 2014. R package version 0.1-10.

[8] J. Feng, S. Jegelka, S. Yan, and T. Darrell. Learning scalable discriminative dictionary with sample relatedness. CVPR, 2014.

[9] Y. Han, F. Wu, X. Lu, Q. Tian, Y. Zhuang, and J. Luo. Correlated attribute transfer with multi-task graph-guided fusion. ACM Multimedia, 2012.

[10] Y. Han, Y. Yang, Z. Ma, H. Shen, N. Sebe, and X. Zhou. Image attribute adaptation. *IEEE Trans. Multimed*, 16(4):1115–1126, 2014.

[11] D. Jayaraman, F. Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*, 2014.

[12] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. CVPR, 2009.

[13] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[14] Q. Qiu, Z. Jiang, and R. Chellappa. Sparse dictionary-based representation and recognition of action attributes. ICCV, 2011.

[15] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. ICML, 2007.

[16] I. Ramirez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. CVPR, 2010.

[17] J. Song, Y. Yang, Z. Huang, H. T. Shen, and R. Hong. Multiple feature hashing for real-time large scale near-duplicate video retrieval. In *ACM MM*, 2011.

[18] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *ACM SIGMOD*, 2013.

[19] S. Wang, X. Chang, X. Li, Q. Z. Sheng, and W. Chen. Multi-task support vector machines for feature selection with shared knowledge discovery. *in press, 2015*.

[20] D. Xu, Y.-L. Chen, C. Lin, X. Kong, and X. Wu. Real-time dynamic gesture recognition system based on depth perception for robot navigation. In *IEEE Int Conf on Robotics*, 2012.

[21] D. Xu, Y.-L. Chen, X. Wu, Y. Ou, and Y. Xu. Integrated approach of skin-color detection and depth information for hand and face localization. In *IEEE Int Conf on Robotics*, 2011.

[22] Y. Yan, G. Liu, E. Ricci, and N. Sebe. Multi-task linear discriminant analysis for multi-view action recognition. In *ICIP*, 2013.

[23] Y. Yan, E. Ricci, G. Liu, and N. Sebe. Recognizing daily activities from first-person videos with multi-task clustering. In *ACCV*, 2014.

[24] Y. Yan, E. Ricci, R. Subramanian, O. Lanz, and N. Sebe. No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *ICCV*, 2013.

[25] Y. Yan, H. Shen, G. Liu, Z. Ma, C. Gao, and N. Sebe. Glocal tells you more: Coupling glocal structural for feature selection with sparsity for image and video classification. *CVIU*, 2014.

[26] Y. Yan, Z. Xu, G. Liu, Z. Ma, and N. Sebe. Glocal structural feature selection with sparsity for multimedia data understanding. In *ACM MM*, 2013.

[27] Y. Yan, Y. Yang, H. Shen, D. Meng, G. Liu, A. Hauptmann, and N. Sebe. Complex event detection via event oriented dictionary learning. In *AAAI*, 2015.

[28] Y. Yang, Z. Ma, Z. Xu, S. Yan, and A. G. Hauptmann. How related exemplars help complex event detection in web videos? In *ICCV*, 2013.

[29] Y. Yang, F. Wu, F. Nie, H. T. Shen, Y. Zhuang, and A. G. Hauptmann. Web and personal image annotation by mining label correlation with relaxed visual graph embedding. *IEEE Transactions on Image Processing*, 21(3):1339–1351, 2012.