Correlated Attribute Transfer with Multi-task Graph-Guided Fusion

Yahong Han
School of Computer Science and Technology
Tianjin University, China
yahong@tju.edu.cn

Fei Wu
College of Computer Science
Zhejiang University, China
wufei@zju.edu.cn

Xinyan Lu
College of Computer Science
Zhejiang University, China
luxinyan1988@gmail.com

Qi Tian
Department of Computer Science
University of Texas at San Antonio
qitian@cs.utsa.edu

Yueting Zhuang
College of Computer Science
Zhejiang University, China
yzhuang@zju.edu.cn

Jiebo Luo
Department of Computer Science
University of Rochester
jluo@cs.rochester.edu

ABSTRACT

Due to the *describable* or *human-nameable* nature of visual attributes, the attribute-based methods have been receiving much attentions in recent years in many applications. The advantages of the utilization of visual attributes are that they can be composed to create descriptions at various levels of specificity or they can be learned once and then applied to recognize new objects or categories. Therefore, attribute prediction becomes an essential problem to boost image understanding. This paper proposes an approach for correlated attribute transfer from a well-defined *source* image set to an uncontrolled *target* image set for attribute prediction. We call it correlated attribute transfer with multi-task graph-guided fusion (CAT-MtG^{2}F). The novelty of CAT-MtG^{2}F is to encourage highly correlated attributes to share a common set of relevant low-level features and transfer the learned common structure from the source image set to the target image set. The experiments show that the proposed CAT-MtG^{2}F achieves better performance in attribute prediction.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

Keywords

Feature Selection, Multi-task Graph-Guided Fusion, Attribute Prediction, Transfer Learning

MM' 12, October 29–November 2, 2012, Nara, Japan.

1. INTRODUCTION

In recent years, the recognition of object-level categories in images has become a major focus of computer vision and multimedia retrieval. Different from the tag-based image understanding [13] and retrieval [33], however, an object also has many other intrinsic human-nameable qualities (*visual attributes*) apart from its corresponding category (semantic tag) and they can be computed to boost image understanding [9]. For example, a zebra can be stripped and a wolf is furry.

As a middle-level cue that bridges low-level image features and high-level object classes, visual attributes have been shown to be beneficial for improving the performance of image understanding [12, 18, 23]. For examples, attribute-based classification is proposed in [18] to facilitate zero-shot learning or serendipitous transfer learning [39] by sharing information via an attribute layer. Different from the prediction of the presence of an attribute in a single image, relative attributes are computed to indicate the strength of an attribute in an image with respect to other images in [23]. It is quite obvious that appropriate learning and inferring visual *attributes* of objects in images provide a compelling way to improve image understanding.

Using the same intuitions in [4, 35] that different features describe different aspects of visual characteristics, Hwang *et al.* [15] assumed that all attributes can rely on some shared structure in their corresponding original low-level visual feature space, and proposed a convex multi-task learning approach with an ℓ_1/ℓ_2-norm to learn a shared low dimensional representation of attributes. The goal is to favor common sparsity patterns in attributes.

However, according to the work in [15], not all attributes will benefit feature sharing, and therefore traditional multi-task learning methods [1, 38, 24] will be detrimental for attribute prediction. These approaches are intended for learning multiple related tasks jointly. The goal is to induce a common-across-tasks sparse representation by analyzing data from all of the tasks at the same time, instead of considering each task individually.

This point in [15] is also supported by the images in the real-worlds. Using the *Animals with Attributes* image dataset (AwA) [18] as an example, AwA contains 30,475 images, 50

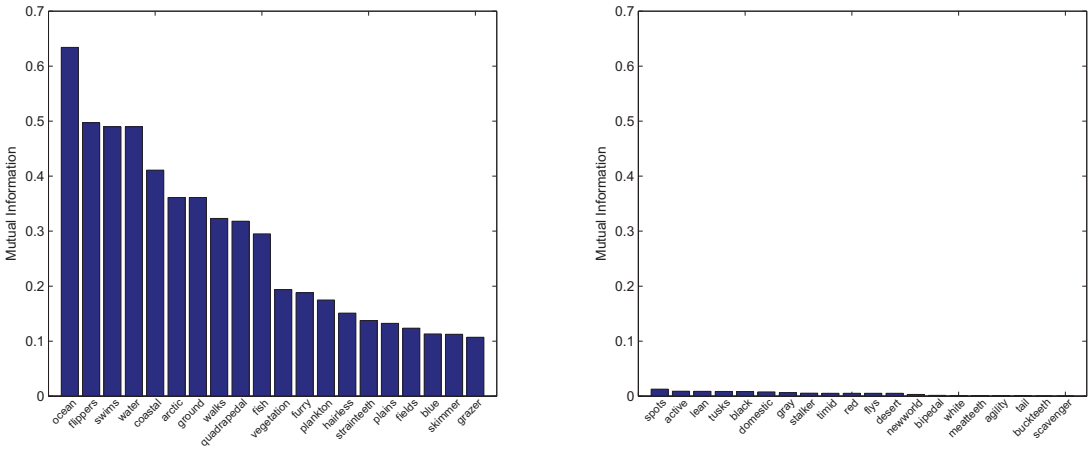

Figure 1: Mutual information of the "ocean" attribute with other attributes in *Animals with Attributes* (AwA) image dataset. Left: 20 attributes in AwA with highest mutual information with attribute "ocean". Right: by contrast, 20 attributes in AwA with lowest mutual information with attribute "ocean".

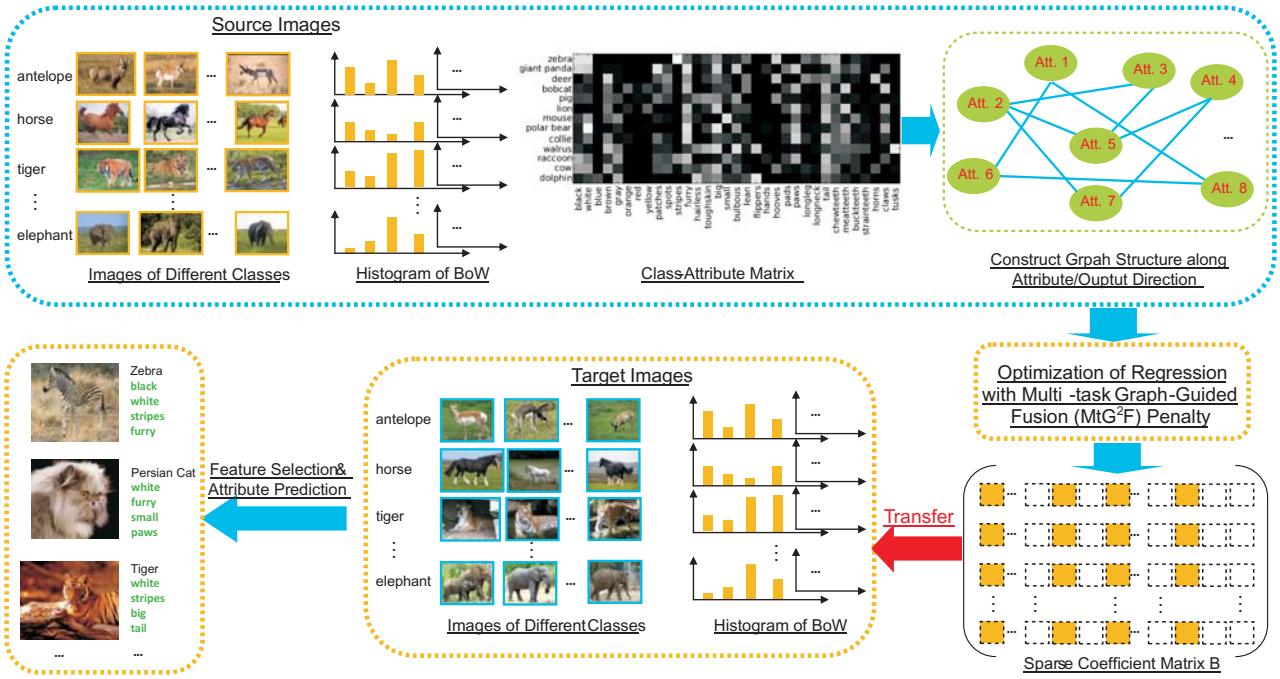

Figure 2: The Flowchart of Correlated Attribute Transfer with Multi-task Graph-Guided Fusion (CAT-MtG^{2}F) penalty. We extracted SIFT features for the source and target images and represent each image by the histogram of Bag of visual Words (BoW). Given the class-attribute matrix, we construct a graph structure along the attribute (output) direction. Then we optimize a regression process regularized with Multi-task Graph-Guided Fusion (CAT-MtG^{2}F) and ℓ_1-norm penalty. The output sparse coefficient matrix can be used to transfer the correlated information among attributes in the source images and conduct feature selection. Finally, we predict attributes for the target images from the selected visual features by using the trained classifiers.

animal classes, and 85 attributes. Each image is labeled by the animal class and related attributes. Figure 1 illustrates the mutual information of the "*ocean*" attribute with other left attributes in AwA. As we can see, the "*ocean*" attribute has different correlations with other attributes, for example, a strong correlation with the "*water*" attribute and a weak correlation with the "*dessert*" attribute. That is to say, given the images in the real-world, some attributes are often more closely related and more likely to share common relevant low-level features than other attributes. Therefore, the introduction of a more appropriate structured regularizer into multi-task learning can discover the shared visual features of related attributes, which poses a great challenge for image understanding.

As a result, unlike the utilization of a mixed-norm regularization such as ℓ_1/ℓ_2- and ℓ_l/ℓ_∞-norms in previous multi-task learning that assumes all tasks (outputs) to be equally related to inputs (features), we investigate a feature selection approach to encourage highly correlated attributes to share a common set of relevant visual features. As sparse coefficient matrix is particular suitable for feature selection [36][35][34], we propose a sparsity-based feature selection method.

Furthermore, since visual attributes are human-specified descriptions of the objects in images, the collection of well-labeled training set for attribute prediction is a tedious work. Assuming that we are given a well-defined *source* image set, in which each image has appropriately labeled attributes and categories, we tend to utilize the cues in the *source* image set to predict the attributes of images in the *target* set.

In order to achieve this goal, this paper is focused on attribute learning by a graph regularizer with structured sparsity in the *source* image set and transferring the learned knowledge to a *target* image set for attribute prediction. We call this correlated attribute transfer with multi-task graph-guided fusion (CAT-MtG^{2}F). The flowchart of CAT-MtG^{2}F is shown in Figure 2. CAT-MtG^{2}F includes three key steps: (1) learning the model of feature selection with correlated attribute transfer (conducted in *Source Images*); (2) feature selection for each attribute (conducted in *Target Images*); (3) predicting attributes from the selected visual features using the trained classifiers.

The strength of our proposed CAT-MtG^{2}F lies in that the correlation information in the output variables (attributes) are naturally exploited. We formulate CAT-MtG^{2}F to be a multi-task learning process so as to discover the shared structures of low-level features for related attributes. CAT-MtG^{2}F uses a graph structure as the guide to find the set of relevant input features that can jointly affect highly correlated outputs. Figure 2 gives out the implementation details of the CAT-MtG^{2}F.

The remainder of this paper is organized as follows. We first introduce the related work in Section 2, and provide the details of CAT-MtG^{2}F in Section 3. In Section 4, we present the construction of an attribute graph. The experimental analysis and conclusions are given in Section 5 and Section 6, respectively.

2. RELATED WORKS

2.1 Attribute Learning for Image Understanding

Due to the *describable* or *human-nameable* nature of visual attributes, attribute-based methods have shown promising potential in object recognition and have recently been receiving much attention. Its success in recognition is largely accredited to the introduction of "attribute", a middle-level cue with meaningful semantical representation. For example, some research works have shown their impact for strengthening face verification [17], object recognition [8], generating natural language descriptions from unfamiliar images [16], facilitation of "zero-shot" transfer learning [18], and prediction of the relative strength of an attribute with respect to other novel images [23].

For the success of aforementioned attribute learning approaches, training of attribute classifiers or regressors from low-level features plays an important role. In these cases, regularization is essential because the inputs (low-level image features) are very high-dimensional with complex variations, and training data are always limited.

In this paper, we are interested in attribute learning by a graph regularizer with structured sparsity in the source image set and transfer the learned knowledge to the *target* image set for attribute prediction. A common goal of transfer learning methods is to discover the common embedded representations from source data and introduce such representation to similar but different target data [26]. In this domain adaptation setting, transfer learning is used to leverage the data available in the source domain to improve the accuracy of the model when testing on target domain examples. The rationale behind transfer learning is to bring the empirical source distribution closer to the target domain, and thus increase the accuracy of the classifier when evaluated on the target domain data. For examples, [18] proposed direct attribute prediction (DAP) and indirect attribute prediction (IAP) for attribute-based classification by transferring information between disjoint training and test classes.

2.2 Structured Sparsity and Fused-lasso based Methods

In the real-world, we can extract high dimensional low-level features from a given image and only limited features distinguish each attribute in the image from others. The problem is that the same visual features might be trivial in one context but crucially significant in another, i.e., some of visual features are most visually important for many related attributes and low-level features are attribute-dependent (or scene-dependent) [28, 15].

As a result, *sparsity* is a reasonable assumption since only a small fraction of visual features are associated with their corresponding attributes. As one of *sparsity*-based feature selection approaches, ℓ_1-norm (namely *lasso*, least absolution shrinkage and selection operator) tends to set up an interpretable model for feature selection [30]. However, *lasso* is prone to selecting features individually and cannot be applied to encode certain structured prior into the process of high-dimensional feature selection.

In order to discover the shared structure of visual features for correlated attributes, it is natural to treat each attribute as a task and perform multi-task learning (MTL) over all of attributes (tasks). The idea of MTL is that, when the tasks to be learned are similar enough or are related in some sense, it may be advantageous to take into account these relations between tasks. In general, MTL can be formulated as a regularization framework [7]. For example, an ℓ_1/ℓ_2-norm is proposed in [1, 20] to learn a common subset of features across all tasks, while the ℓ_l/ℓ_∞-norm is used to promote

feature sharing across tasks and discover solutions where only a few features are non-zero in any of the tasks (i.e. jointly sparse solutions) [25].

Graph has been successfully utilized to model the context correlations of images [29, 13] and videos [32]. For example, different relationships among video units are well represented and unified by multiple graphs [32]. To explicitly incorporate the graph-structured context constraint into the sparse learning framework, some previous structured sparse learning methods could partially accomplish this task. The fused lasso [31] penalizes both the ℓ_1-norm of the coefficients and their successive differences. The grouping pursuit [27] utilizes a penalty involving pairwise comparisons for each pair of coefficients. That is, the grouping pursuit assumes each pair of variables are all correlated. It is worth noting that, the penalties by fused lasso and grouping pursuit introduce two special graph structures, i.e., a chain of successive nodes and a full-connected graph of all the nodes, respectively. To be more general than the fused lasso and grouping pursuit, we need a structured sparse penalty defined on the general graph structure so as to incorporate the graph-structured context constraint into the sparse reconstruction framework.

According to [15], during the learning of shared features of attributes, some of the attributes are often more closely related and more likely to share common relevant visual features than other attributes. Thus, it is desirable to take into account the complex correlation structure in the outputs for a more effective multi-task learning. Chen *et al.* [6] proposed a structured regularized regression approach called graph-guided fused lasso (GFlasso) for sparse multi-task learning problems to encourage tasks correlated in the graph to have a similar sparsity pattern. GFlasso employs a "fusion penalty" [31] that fuses regression coefficients across correlated outputs by using the weighted connectivity of the output graph as a guide.

2.3 Transfer Learning and Feature Representation Transfer

Transfer learning aims to help improve the learning of the target predictive function in a target domain using the knowledge in a source domain [22]. In multimedia and computer vision research, many works utilized knowledge transfer to improve the performance [14, 26, 37]. In [26], sparse prototype representations were learned from source images, which are used to get better performance in target images. Similarly, in [14], a multi-label encoded sparse linear embedding space was learned from source domain. Mapping target images into the learned new representations, better performance of image annotation can be obtained. According to the survey in [22], transferred approaches in [14] and [26] can be categorized into the "feature-representation-transfer", which is similar in spirit as the proposed framework of CAT-MtG^{2}F. In transferred framework of CAT-MtG^{2}F, see Fig. 2, the correlated information among attributes in source images is encoded into a learned sparse coefficient matrix, in which each column represents correlated visual features for each attribute and thus can be used to perform feature selection for target images. Thus, the correlated information among attributes is transferred and the goal is to get a better representation for target images. Though knowledge transfer is also used to perform cross-view action recognition in [19], the transferred knowledge of [19] is between two views of video cameras, which is different from the proposed framework of CAT-MtG^{2}F.

3. FEATURE SELECTION WITH CORRELATED ATTRIBUTE TRANSFER

3.1 Notation

Assume that we have a training set of N labeled images from J semantic categories. Matrix $\mathbf{X} \in \mathbb{R}^{N\times P}$ represents the image data matrix, P is the dimensionality of image features. Matrix $\mathbf{Y} \in \{0,1\}^{N\times J}$ represents the label indicator matrix of image data, $\mathbf{Y}_{(i,j)} = 1$ if the ith image belongs to the jth categories and $\mathbf{Y}_{(i,j)} = 0$ otherwise. Let $\mathbf{A} \in \{0,1\}^{J\times L}$ denote the class-attribute matrix, in which $\mathbf{A}_{(j,l)} = 1$ if the jth class has the lth attribute and $\mathbf{A}_{(j,l)} = 0$ otherwise. Given matrices $\mathbf{Y}$ and $\mathbf{A}$, we can construct matrix $\mathbf{T} \in \{0,1\}^{N\times L}$ to represent the attribute indicator matrix of image data, $\mathbf{T}_{(i,l)} = 1$ if the ith image has the lth attributes and $\mathbf{T}_{(i,l)} = 0$ otherwise. Matrix $\mathbf{B} \in \mathbb{R}^{P\times L}$ represents the coefficient matrix, and $\mathbf{B}_{(:,l)}$ $(l = 1, \ldots, L)$ is the coefficient vector for the lth attribute. We let $||\mathbf{M}||_F$ denote the matrix Frobenius norm, $||\mathbf{M}||_1$ denote the matrix entry-wise ℓ_1-norm, $\mathbf{x}^T$ and $\mathbf{M}^T$ represent the transpose of vector $\mathbf{x}$ and matrix $\mathbf{M}$, respectively.

Without loss of generality, $\mathcal{S} = \{\mathbf{X}^{\mathcal{S}}, \mathbf{Y}^{\mathcal{S}}, \mathbf{A}^{\mathcal{S}}\}$ denotes the visual features, label indicator and class-attribute from source image dataset $\mathcal{S}$ and are available during training. $\mathcal{T} = \{\mathbf{X}^{\mathcal{T}}, \mathbf{Y}^{\mathcal{T}}, \mathbf{A}^{\mathcal{T}}\}$ denotes the visual features, label indicator and class-attribute from target image dataset $\mathcal{T}$ and only $\mathbf{X}^{\mathcal{T}}$ is available during testing. The goal of this paper is to boost the performance of predicting the attribute indicator matrix $\mathbf{T}^{\mathcal{T}}$ of target images by borrowing strength from the source image dataset $\mathcal{S}$.

Algorithm 1 Feature Selection with Correlated Attribute Transfer (FSCAT)

Input: source image dataset $\mathcal{S} = \{\mathbf{X}^{\mathcal{S}}, \mathbf{Y}^{\mathcal{S}}, \mathbf{A}^{\mathcal{S}}\}$; target images $\mathcal{T} = \{\mathbf{X}^{\mathcal{T}}\}$; k is the number in the k nearest neighbor graph construction; f is the number of features to be selected.

Output: index ind_l of the top f selected features, $l = 1, \ldots, L$

1: Construct graph G on columns of matrix $\mathbf{A}$ using the mutual k nearest neighbor (kNN) strategy;
2: Solve Eq. (1) and output the coefficient matrix $\hat{\mathbf{B}}$;
3: For $l = 1, \ldots, L$
4: Rank elements in vector $\mathbf{B}_{(:,l)}$ according to the value of $|\mathbf{B}_{(p,l)}|$ $(p = 1, \ldots, P)$ in descending order;
5: Output the index ind_l of the top f selected features;
6: End For

3.2 Learning Correlated Attributes from Source Images

Suppose we have a well-devised source image dataset $\mathcal{S} = \{\mathbf{X}^{\mathcal{S}}, \mathbf{Y}^{\mathcal{S}}, \mathbf{A}^{\mathcal{S}}\}$, our goal is to predict $\mathbf{T}^{\mathcal{T}}$ of the target images $\mathcal{T} = \{\mathbf{X}^{\mathcal{T}}\}$. As introduced above, we can construct matrix $\mathbf{T}^{\mathcal{S}} \in \{0,1\}^{N\times L}$ from $\mathbf{Y}^{\mathcal{S}}$ and $\mathbf{A}^{\mathcal{S}}$ to represent the attribute indicator matrix of source image data. Since one class may correlate multiple attributes, correlations among attributes are encoded in matrix $\mathbf{T}^{\mathcal{S}}$. In order to utilize the correlated

information in the source images, we formulate a multi-task learning process and learn the the coefficient matrix $\mathbf{B}$ for transferred feature selection in the target images.

We propose to optimize the following multi-task regression problem:

$$\min_{\mathbf{B}\in\mathbb{R}^{P\times L}} ||\mathbf{T}^{\mathcal{S}} - \mathbf{X}^{\mathcal{S}}\mathbf{B}||_F^2 + \gamma\Omega_G(\mathbf{B}) + \lambda||\mathbf{B}||_1, \qquad (1)$$

where $\Omega_G(\mathbf{B})$ is a structured sparsity-inducing norm penalty with a structure G defined over the output direction of $\mathbf{B}$. That is to say, G is defined over L attributes to represent specific correlation structure. $||\mathbf{B}||_1$ is the entry-wise matrix ℓ_1-norm of B (i.e.,$||\mathbf{B}||_1 = \sum^P\sum^L|\mathbf{B}(:,:)|$). γ and λ are regularization parameters respectively that control the complexity of the model. A larger value of γ leads to a greater fusion effect. Assuming that a graph structure over the L attributes is given as G with a set of nodes $V = \{1,\ldots,L\}$ each corresponding to an attribute and a set of edges E. Thus the Multi-task Graph-Guided Fusion (MtG^{2}F) penalty for a structured multi-task regression is given as:

$$\Omega_G(\mathbf{B}) = \sum_{e=(m,l)\in E, m<l} w_{ml}\sum_{p=1}^{P}|\mathbf{B}_{(p,m)} - \mathbf{B}_{(p,l)}|, \qquad (2)$$

where $\mathbf{B}_{(p,m)}$ and $\mathbf{B}_{(p,l)}$ are coefficients in $\mathbf{B}$ corresponding to the coefficients of the mth and lth attributes for the pth feature, respectively. The weight w_{ml} measures the fusion penalty for each edge $e = (m,l)$ such that $\mathbf{B}_{(p,m)}$ and $\mathbf{B}_{(p,l)}$ for highly correlated attributes with larger w_{ml} receive a greater fusion effect. Therefore, the multi-task graph-guided fusion (MtG^{2}F) penalty in Eq. (1) encourages highly correlated attributes corresponding to a densely connected subnetwork in G to be jointly encoded in the multi-task regression process. Furthermore, the matrix entry-wise ℓ_1-norm $||\mathbf{B}||_1$ in Eq. (1) encourage sparsity in matrix $\mathbf{B}$, making it particularly suitable for feature selection. If we construct the graph structure G appropriately, e.g., encoding correlation information among the attributes into the graph structure, we can expect a better coefficient matrix $\mathbf{B}$, which can be used to transfer the information of correlated attributes and perform feature selection.

3.3 Selecting Transferred Features in Target Images

Let the estimated coefficient matrix $\hat{\mathbf{B}}$ be the output of solving Eq. (1), many coefficients in $\hat{\mathbf{B}}$ are expected to shrink to zeros (both in rows and columns) owing to the penalty of the matrix entry-wise ℓ_1-norm $||\mathbf{B}||_1$ in Eq. (1). In other words, for predicting the mth attribute, the learned discriminative regression coefficient vector $\mathbf{B}_{(:,m)}$ only utilizes the features when the corresponding elements in vector $\mathbf{B}_{(:,m)}$ are not zero. Alternatively, we can rank each feature of the ith target images $\mathbf{X}^{\mathcal{T}}_{(i,p)}|_{p=1}^{P}$ according to the value of $|\mathbf{B}_{(p,m)}|$ in descending order and select top ranked features. In the following, we summarize the algorithm of Feature Selection with Correlated Attribute Transfer in Algorithm 1

3.4 Solutions and Algorithm

In this section, we firstly introduce the solutions of structured multi-task regression optimization for attribute transfer in Eq. (1), then we will summarize the algorithm of attribute prediction for target images.

3.4.1 Optimization of Structured Multi-task Regression for Attribute Transfer

Although the optimization problem in Eq. (1) is convex, the main difficulty for the optimization of Eq. (1) arises from the non-separability of $\mathbf{B}$ in the non-smooth penalty function $\Omega_G(\mathbf{B})$. In this paper, we solve Eq. (1) by a general smoothing proximal gradient (SPG) method [6]. The SPG first finds a separable and smooth approximation of $\Omega_G(\mathbf{B})$, and then solves this transformed simple ℓ_1-norm penalized sparse learning problem by the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [2]. According to SPG [6], key steps of the approximation process are as follows.

(1) The Multi-task Graph-Guided Fusion (MtG^{2}F) penalty $\Omega_G(\mathbf{B})$ is reformulated into a linear transformation of $\mathbf{B}$ via the dual norm as follows:

$$\Omega_G(\mathbf{B}) = \max_{\mathbf{A}\in\mathcal{Q}} < C\mathbf{B}^T, \mathbf{A} >, \qquad (3)$$

where $< U, V >\equiv \mathrm{Tr}(U^TV)$ denotes a matrix inter product. $\boldsymbol{\alpha} \in \mathcal{Q} = \{\boldsymbol{\alpha}|||\boldsymbol{\alpha}||_\infty \le 1, \boldsymbol{\alpha} \in \mathbb{R}^{|E|}\}$ is a column vector in the auxiliary matrix $\mathbf{A}$ and is associated with $||C\mathbf{B}^T||_1$. Matrix $C \in \mathbb{R}^{|E|\times L}$ is the edge-vertex incident matrix with the form of

$$C_{e=(m,l),l'} = \begin{cases} w_{ml}, & \text{if } l' = m, \\ -w_{ml}, & \text{if } l' = l, \\ 0, & \text{otherwise,} \end{cases} \qquad (4)$$

where $l' = 1,\ldots,L$ index the output (i.e., the attribute) direction of matrix $\mathbf{B}$.

(2) Introduce the smoothing approximation function $f_\mu(\mathbf{B})$ as follows:

$$f_\mu(\mathbf{B}) = \max_{\mathbf{A}\in\mathcal{Q}} < C\mathbf{B}^T, \mathbf{A} > -\mu d(\mathbf{A}), \qquad (5)$$

where $\mu \ge 0$ is the smoothness parameter and $d(\mathbf{A}) \equiv \frac{1}{2}||\mathbf{A}||_F^2$.

Substitute $\Omega_G(\mathbf{B})$ in Eq (1) with $f_\mu(\mathbf{B})$, our goal is to solve the following optimization problem:

$$\min_{\mathbf{B}\in\mathbb{R}^{p\times L}} ||\mathbf{T}^{\mathcal{S}} - \mathbf{X}^{\mathcal{S}}\mathbf{B}||_F^2 + \gamma f_\mu(\mathbf{B}) + \lambda||\mathbf{B}||_1, \qquad (6)$$

Because the first two terms of Eq. (6) are convex and smooth, Eq. (6) can be efficiently solved by the FISTA algorithm [2]. It has been proved in [6] that, SPG can achieve $O(\frac{1}{\epsilon})$ convergence rate for a desired accuracy ϵ.

3.4.2 Attribute Prediction for Target Images

In this section, we will introduce the algorithm of attribute prediction for target images. Specifically, the algorithm includes three key steps: (1) learning the model of feature selection with correlated attribute transfer (conducted in *Source Images*); (2) feature selection for each attribute (conducted in *Target Images*); (3) predict attributes from the selected visual features by the trained classifiers, e.g., the kNN classifier. We summarize the algorithm of attribute prediction for target images in Algorithm 2.

Thus, we propose the Correlated Attribute Transfer with Multi-task Graph-Guided Fusion (CAT-MtG^{2}F) in Algorithm 1 and 2. Since we have the groundtruth of class-attribute matrix $\mathbf{A}$, see Section 3.1, we can straightforwardly conduct class detection for testing images in target dataset after we successfully predict the attributes of the images

Algorithm 2 Attribute Prediction for Target Images

Input: source image dataset $\mathcal{S} = \{\mathbf{X}^{\mathcal{S}}, \mathbf{Y}^{\mathcal{S}}, \mathbf{A}^{\mathcal{S}}\}$; target images $\mathcal{T} = \{\mathbf{X}^{\mathcal{T}}\}$; k' is the number in the k'NN classifier.
Output: attribute indicator matrix $\mathbf{T}^{\mathcal{T}}$ for testing images of target data

1: Output the index ind_l $(l = 1, \ldots, L)$ of visual feature selection for each attribute by FSCAT in Algorithm 1;
2: For $l = 1, \ldots, L$
3: Perform feature selection of target images by

$$\mathbf{X}^T_{selected} = \mathbf{X}^{\mathcal{T}}(:, ind_l);$$

4: Predict attribute by training and testing on $\mathbf{X}^T_{selected}$ using k'NN classifier;
5: End For
6: Output the attribute indicator matrix $\mathbf{T}^{\mathcal{T}}$ for testing images in the target dataset.

by Algorithm 2. Therefore, the proposed framework of feature selection with correlated attribute transfer for image understanding dedicates to predict both attribute and corresponding semantic category for target images.

4. ATTRIBUTE GRAPH CONSTRUCTION

One semantic category may correlate multiple attributes, e.g., attributes "black", "white", and "stripes" are taken as the high-level description of image semantic category "zebra". On the other hand, several correlated attributes are usually used simultaneously to describe the multiple semantic categories, e.g., attributes "water" and "eat fish" appear to describe image class "otter" and "polar bear". Taking the class-attribute matrix $\mathbf{A} \in \{0, 1\}^{J \times L}$ in the AwA dataset [18] (as shown in Figure 2) for example, let each column vector represent one attribute and then the attributes with similar binary patterns in the corresponding column vectors will be correlated with each other.

We use the mutual k nearest neighbor (kNN) strategy to discover the relationship between attributes. Specifically, two attributes (represented by column vectors $\mathbf{A}_{(:,m)}$ and $\mathbf{A}_{(:,l)}$ in $\mathbf{A}$) are linked with an edge $e = (m, l) \in E$ if $\mathbf{A}_{(:,m)}$ is among the k nearest neighbors of $\mathbf{A}_{(:,l)}$ and vice versa. In this setting, graph G is undirected and the adjacent relations between two regions are symmetric.

Euclidean distance is used to calculate the graph weight w_{ml} as follows:

$$w_{ml} = \begin{cases} e^{-\frac{||\mathbf{A}_{(:,m)} - \mathbf{A}_{(:,l)}||_2^2}{2\sigma^2}}, & \text{if } m \in N_k(l) \text{ or } l \in N_k(m), \\ 0, & \text{otherwise}, \end{cases} \quad (7)$$

where $N_k(m)$ and $N_k(l)$ denote the vertex index set for the k nearest neighbors of the mth and lth vertex, respectively.

5. EXPERIMENTS

In this section, we evaluate the effectiveness of our proposed CAT-MtG^{2}F algorithm for the task of attribute learning for target images. To show the competitive performance of transferred feature selection for attribute learning, we compare our algorithm with other state-of-the-art feature selection methods. The experimental results show that our algorithm achieves a better performance of attribute prediction owing to the strength transferred from the source images. Furthermore, the sparse coefficient matrix $\hat{\mathbf{B}}$ (see Eq. (1)) are obtained by introducing the graph structure among attributes into the process of multi-task regression. In this experiment, we will also demonstrate that the learned matrix $\hat{\mathbf{B}}$ successfully discovers some latent correlations between attributes and visual features (the SIFT BoW used in this experiment).

(a) Images selected from AwA (source) image dataset

(b) Images selected from AIN (target) dataset

Figure 3: Sample images selected from the source and target image datasets. Images in the top rows are from the category "dolphin". Images in the bottom rows are from the category "fox".

5.1 Dataset and Feature Extraction

5.1.1 *Source Images and Target Images*

Two image datasets are respectively used as source images and target images in this experiment. A well-devised "Animals with Attributes" (AwA) dataset [18] is used as source images, which contains 30,475 animal images that are categorized into 50 classes. To boost the research of attribute based classification [18] and image understanding with attribute [18, 8, 23], Lampert *et al.* align the 50 animals classes with the classical class-attribute matrix [21], thus providing 85 numeric attribute values for each class, which can be taken as the groundtruth of class-attribute correlations. Images in each class of AwA are crawled and selected from the Web. So AwA can be considered as a well-devised source image datase.

For target images, we submitted the 50 animal (class) names in AwA as 50 query keywords to ImageNet and downloaded images for each class. As a result, we collected altogether 48,199 images to form the target image dataset and these images are categorized into the same animal classes as in AwA. Thus, the collected images are appropriate to

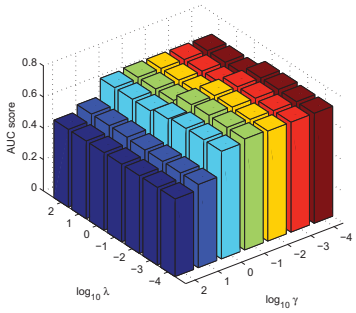

Figure 4: Parameter tuning results of γ and λ. The ranges of γ and λ are {0.0001, 0.001, 0.01, 0.1, 1, 10, 100}

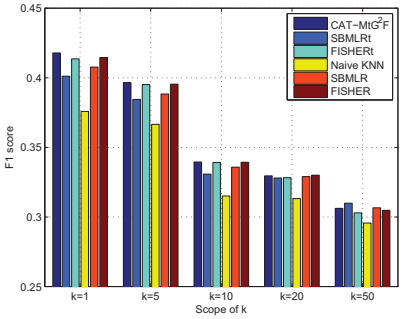

Figure 5: Performance comparison when k is set to different values. The results were obtained when 10% of target images are used as training images.

be considered as the target images and used to evaluate the performance of the proposed CAT-MtG^{2}F algorithm. In the following, we denote this dataset as "Animals from ImageNet" (AIN). To demonstrate the differences between the source images and target images, we show some sample images in Figure 3(a) and Figure 3(b).

Moreover, to further demonstrate the ability of attribute prediction of CAT-MtG^{2}F, we also transfer knowledge learned from AwA to images in the dataset of "Animal on the Web" (AoW) [3]. Thus, two target image datasets AIN and AoW are respectively used in this experiment to be the test bed for evaluation of the correlated attribute transfer from source images in AwA.

5.1.2 Visual Features

As shown in Figure 3, though coming from the same semantic category, images may have very variant visual features, e.g., variant colors, variant shapes, variant textures etc. Inspired by the research in object recognition and detection [11], the keypoint based local features are more robust. In this experiment, we first extract SIFT points for each images in AwA, AIN, and AoW. The randomly selected set of SIFT points are clustered and produced the 1,000 centers as the visual dictionary. Then each image is quantized into a 1,000 dimensional histogram of bag-of-visual-words (BoW).

5.1.3 Partition of Training/Testing Images

For all the 48,199 images in AIN dataset, we randomly sample 10,000 images to form the target images for algorithm evaluation. For the partition of training/testing images in the selected target image dataset, we randomly sample 1,000 to 3,000 samples to form the training data and the remaining data are used as the corresponding testing images. For each partition, all above sampling processes are repeated ten times to generate 10 random training/test partitions. The average performance and standard deviation are reported.

5.2 Evaluation Metrics

The F_1 score is used to measure the performance of attribute prediction, which is defined as the harmonic mean of precision and recall:

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

Since there are multiple attributes in our experiment, we use the macroaveraging scheme and average the F_1 scores over multiple attributes.

In the process of parameter (γ and λ in Eq. (1)) tuning in source image dataset, prediction performance of the learned matrix $\mathbf{B}$ need to be evaluated. Since $\mathbf{B}$ is trained and learned from a structured multi-task regression process, the values of predicted responses are continuous but binary. Thus, in the process of parameter tuning, we use the Aera Under the Roc Curve (AUC) [10] as the evaluation metic.

5.3 Parameter Tuning

Effects of graph-guided fusion and sparsity by ℓ_1-norm penalty is mainly determined by parameters γ and λ in Eq. (1). In this experiment, we use a 5-fold cross validation on the source image dataset to tune the value of parameters γ and λ. We choose those parameters which perform best at the source test dataset and then the values are then fixed throughout the experiments.

The ranges of γ and λ are {0.0001, 0.001, 0.01, 0.1, 1, 10, 100}. The parameter tuning results are shown in Figure 4. We can observe that the highest performance (AUC=0.711) of two parameters is achieved at some intermediate values. Thus we set $\gamma = 0.01$ and $\lambda = 0.1$ and under this parameter setting the ratio of sparsity of matrix $\mathbf{B}$ is 83%.

5.4 Experimental Results

5.4.1 Comparison Methods

Having learned the coefficient matrix $\mathbf{B}$ from source images in AwA dataset by Algorithm 1, we conduct attribute prediction by Algorithm 2 on the target image dataset by feature selection with correlated attribute transfer from source images to target images.

We compare our method CAT-MtG^{2}F with Fisher Score and Sparse Bayesian Multinomial Logistic Regression (denoted by SBMLR in the following) [5]. SBMLR is a sparse multinomial logistic regression method, in which the sparsity arises from the use of a Laplacian prior. Fisher Score is a statistical score which can output feature rankings and is dependent on the choice of the classifier. One can use Fisher's criterion to calculate the Fisher Score of each feature. The larger the Fisher Score is, the more likely this feature is discriminative. Therefore, we perform the feature selection by ranking the scores.

We have two versions of settings for the SBMLR and

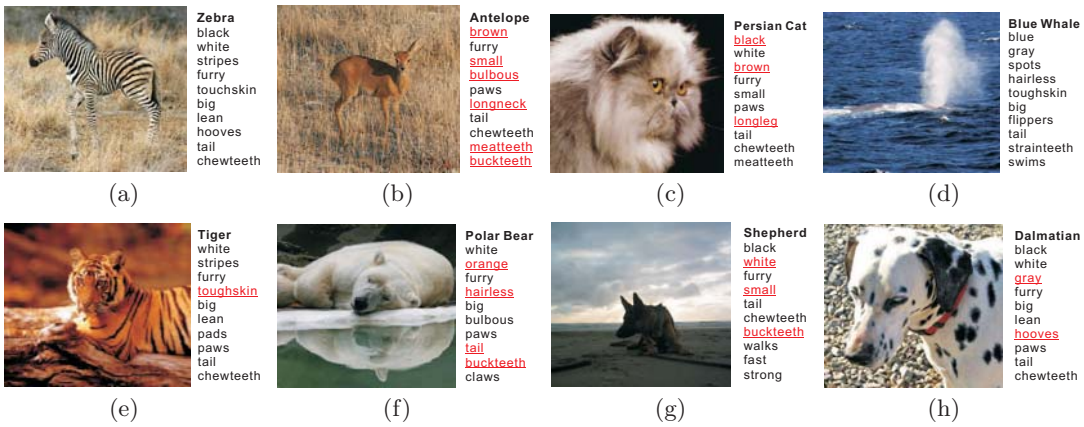

Figure 6: Attribute predictions on the target set (underlined red texts denote the false positive predictions)

Fisher Score algorithms: transferred and without transferr. For the transferred SBMLR and transferred Fisher Score (denoted by SBMLRt and FISHERt for short), algorithms SBMLR and Fisher Score are performed after we conduct feature selection by Step 3 in Algorithm 2, i.e., $\mathbf{X}^T_{selected} = \mathbf{X}^{\mathcal{T}}(:, ind_l)$;. For SMLR and Fisher Score without transfer (denoted by SBMLR and FISHER for short), the algorithms are performed on the original visual features. For all methods, we use kNN classifier to train and predict each binary attribute. To further show the effect of correlated attribute transfer, we also compare CAT-MtG^{2}F with performing kNN on the original visual features of the target images, which is denoted by "Naive kNN" in the following.

5.4.2 Attribute Prediction Results

We first investigate the performance of comparing CAT-MtG^{2}F with state-of-the-art methods for k in kNN classifier is set to be different values. In Figure 5, we plot the attribute prediction results from different methods when $k \in \{1, 5, 10, 20, 50\}$. As can be seen, as the value of k increases, the performance of all the methods decreases. Though the performance of CAT-MtG^{2}F is worse than Fisher Score and SBMLR when $k = 50$, the performance of CAT-MtG^{2}F is better than other methods when $k = 1, 5, 10, 20$. Especially, the improvement of CAT-MtG^{2}F is prominent when $k = 1$.

To investigate the performance of attribute prediction with a different number of selected features and different number of training data for training the kNN classifier, we compare and list the results in Table 1 and Table 2. From the results we can observe that: (1) generally, CAT-MtG^{2}F produces better performance of attribute prediction owing to the feature selection with correlated attribute transfer; (2) though not so significant improvement comparing with SBMLR and Fisher Score, the improvement of CAT-MtG^{2}F comparing with Naive kNN is over 5%. Thus the results demonstrate that the effectiveness of the proposed transferred feature selection framework.

In Figure 6, we show some examples of attribute prediction on the AIN dataset. We can see the prediction on "Zebra" matches the truth perfectly. On the other hand, the attribute prediction performances on different animals may differ. To further show the effectiveness of transferred attribute prediction, in Figure 7, we show some examples of attribute prediction for "unseen" animals of AoW dataset.

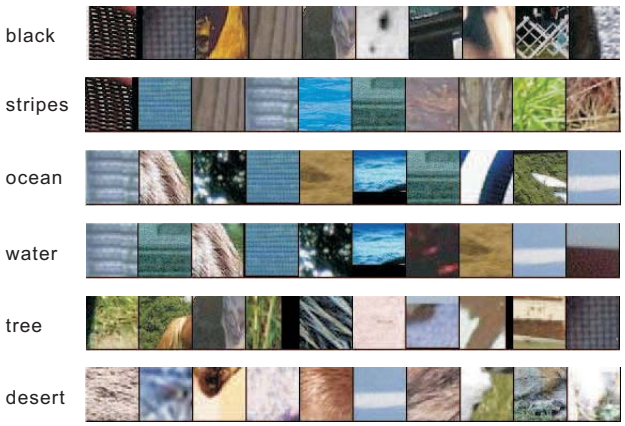

Figure 8: Attribute - Visual words correlation example. Of the 6 example attributes, "ocean" and "water" share high mutual information, while "tree" and "desert" low mutual information. With our method, attribute "ocean" and "water" share the same correlated visual words.

5.4.3 Demonstration of Visual Words and Attributes

Now we look into the matrix $\mathbf{B}$ obtained by solving the regression problem in Eq. (1) . Each column of $\mathbf{B}$ represents correlated feature for each attribute, in which higher amplitude means stronger correlation between attribute and corresponding feature while zero means no correlation. We depict the top 10 highest-ranked visual features (visual words in our experiment) of 6 example attributes in Figure 8.

In human cognition, "ocean" and "water" should be correlated (we have seen that from Figure 1) and they should share similar visual feature. The examples in Figure 8 demonstrates this: "ocean" and "water" have 8 out of the 10 highest-ranked visual words in common. In contrast, "desert" should be an "cognitive" antonym of "tree" and they share no common features: the "tree" attribute's most correlated visual words represents some kind of "energy" while the "desert" attribute represents desolation.

6. CONCLUSIONS

In this work, we have shown how to predict attributes

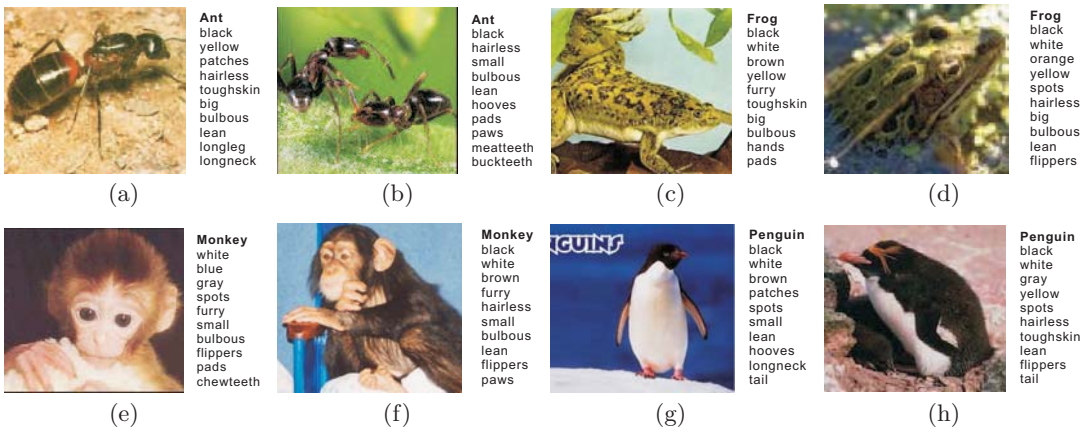

Figure 7: Attribute predictions on AoW dataset

Table 1: Performance comparison with 50 selected features and 10%, 20%, 30% train data separately.

Method / % train data	10%	20%	30%
CAT-MtG^{2}F	**0.4179±0.0043**	**0.4258±0.0025**	**0.4281±0.0017**
SBMLRt	0.4012±0.0028	0.4013±0.0038	0.4024±0.0033
FISHERt	0.4136±0.0042	0.4217±0.0024	0.4254±0.0033
Naive kNN	0.3759±0.0236	0.4033±0.0113	0.4092±0.0092
SBMLR	0.4077±0.0208	0.4154±0.0101	0.4148±0.0050
FISHER	0.4146±0.0033	0.4221±0.0033	0.4239±0.0039

of image objects through the correlated-attribute transfer to boost the image understanding . The multi-task graph-guided fusion penalty utilized in CAT-MtG^{2}F encourages the attributes within a densely connected subgraph to share common low-level features. We demonstrated the effectiveness of the proposed CAT-MtG^{2}F method and showed that our proposed CAT-MtG^{2}F method is comparable to or better than the state-of-the-art methods. Investigation of the correlation between attributes and selected visual words also demonstrated the effectiveness of the proposed transferred feature selection method.

7. ACKNOWLEDGMENTS

This work was supported in part by National Basic Research Program of China (2010CB327904), National HeGaoJi Key Project (2010ZX01042-002-003), NSFC (No. 90920303, 61070068) and the Fundamental Research Funds for the Central Universities. The work of Qi Tian was supported in part by ARO grant W911NF-12-1-0057, NSF IIS 1052851, Faculty Research Awards by Google, NEC Laboratories of America and FXPAL, UTSA START-R award, respectively.

8. REFERENCES

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[2] 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.

[3] T. Berg and D. Forsyth. Animals on the web. In *Proceedings of CVPR 2006*, volume 2, pages 1463–1470. IEEE, 2006.

[4] L. Cao, J. Luo, F. Liang, and T. Huang. Heterogeneous feature machines for visual recognition. In *Proceedings of ICCV 2009*, pages 1095–1102. IEEE, 2009.

[5] G. Cawley, N. Talbot, and M. Girolami. Sparse multinomial logistic regression via bayesian l1 regularisation. *Advances in neural information processing systems (NIPS)*, 19:209, 2007.

[6] X. Chen, Q. Lin, S. Kim, J. Carbonell, and E. Xing. Smoothing proximal gradient method for general structured sparse learning. In *Uncertainty in Artificial Intelligence (UAI)*, 2011.

[7] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD*, pages 109–117. ACM, 2004.

[8] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *Proceedings of CVPR 2010*, pages 2352–2359. IEEE, 2010.

[9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proceedings of CVPR 2009*, pages 1778–1785. IEEE, 2009.

[10] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[11] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[12] V. Ferrari and A. Zisserman. Learning visual attributes. In *Proceedings of Advances in Neural Information Processing Systems (NIPS 2007)*, 2008.

[13] Y. Han, F. Wu, Q. Tian, and Y. Zhuang.

Table 2: Performance comparison with 100 selected features and 10%, 20%, 30% train data separately.

Method / % train data	10%	20%	30%
CAT-MtG^{2}F	**0.4155±0.0063**	**0.4261±0.0045**	**0.4288±0.0027**
SBMLRt	0.4012±0.0028	0.4013±0.0038	0.4024±0.0033
FISHERt	0.4114±0.0059	0.4234±0.0037	0.4271±0.0056
Naive kNN	0.3759±0.0236	0.4033±0.0113	0.4092±0.0092
SBMLR	0.4075±0.0208	0.4154±0.0101	0.4148±0.0050
FISHER	0.4136±0.0047	0.4253±0.0048	0.4274±0.0036

Graph-guided sparse reconstruction for region tagging. In *Proceedings of CVPR 2012*. IEEE, 2012.

[14] Y. Han, F. Wu, Y. Zhuang, and X. He. Multi-label transfer learning with sparse representation. *Circuits and Systems for Video Technology, IEEE Transactions on*, 20(8):1110–1121, 2010.

[15] S. Hwang, F. Sha, and K. Grauman. Sharing features between objects and their attributes. In *Proceedings of CVPR 2011*, pages 1761–1768. IEEE, 2011.

[16] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. Berg, and T. Berg. Baby talk: Understanding and generating simple image descriptions. In *Proceedings of CVPR 2011*, pages 1601–1608. IEEE, 2011.

[17] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (99):1–1, 2011.

[18] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of CVPR 2009*, pages 951–958. IEEE, 2009.

[19] J. Liu, M. Shah, B. Kuipers, and S. Savarese. Cross-view action recognition via view knowledge transfer. In *Proceedings of CVPR 2011*, pages 3209–3216. IEEE, 2011.

[20] G. Obozinski, B. Taskar, and M. I. Jordan. Multi-task feature selection. *Technical Report of UC Berkeley*, 2006.

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

[22] S. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

[23] D. Parikh and K. Grauman. Relative attributes. In *Proceedings of ICCV 2011*. IEEE, 2011.

[24] N. Quadrianto, A. Smola, T. Caetano, S. Vishwanathan, and J. Petterson. Multitask learning without label correspondences. In *Proceedings of Advances in Neural Information Processing Systems (NIPS 2011)*, 2011.

[25] A. Quattoni, X. Carreras, M. Collins, and T. Darrell. An efficient projection for $l_{1,\infty}$ regularization. In *Proceedings of ICML 2009*, pages 857–864. ACM, 2009.

[26] A. Quattoni, M. Collins, and T. Darrell. Transfer learning for image classification with sparse prototype representations. In *Proceedings of CVPR 2008*, pages 1–8. IEEE, 2008.

[27] X. Shen and H. Huang. Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association*, 105(490):727–739, 2010.

[28] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. Efros. Data-driven visual similarity for cross-domain image matching. *ACM Transaction of Graphics (TOG)(Proceedings of ACM SIGGRAPH ASIA)*, 30(6), 2011.

[29] J. Tang, R. Hong, S. Yan, T. Chua, G. Qi, and R. Jain. Image annotation by k nn-sparse graph-based label propagation over noisily tagged web images. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(2):14, 2011.

[30] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[31] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[32] M. Wang, X. Hua, R. Hong, J. Tang, G. Qi, and Y. Song. Unified video annotation via multigraph learning. *Circuits and Systems for Video Technology, IEEE Transactions on*, 19(5):733–746, 2009.

[33] M. Wang, K. Yang, X. Hua, and H. Zhang. Towards a relevant and diverse search of social images. *Multimedia, IEEE Transactions on*, 12(8):829–842, 2010.

[34] F. Wu, Y. Han, X. Liu, J. Shao, Y. Zhuang, and Z. Zhang. The heterogeneous feature selection with structural sparsity for multimedia annotation and hashing: a survey. *International Journal of Multimedia Information Retrieval*, pages 1–13, 2012.

[35] F. Wu, Y. Han, Q. Tian, and Y. Zhuang. Multi-label boosting for image annotation by structural grouping sparsity. In *Proceedings of ACM Multimedia 2010*, pages 15–24. ACM, 2010.

[36] Y. Yang, H. Shen, Z. Ma, Z. Huang, and X. Zhou. $\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *Proceedings of IJCAI 2011)*, pages 1589–1594, 2011.

[37] Y. Yang, Y. Yang, Z. Huang, and H. Shen. Transfer tagging from image to video. In *Proceedings of ACM Multimedia 2011*, pages 1137–1140. ACM, 2011.

[38] K. Yu, V. Tresp, and A. Schwaighofer. Learning gaussian processes from multiple tasks. In *Proceedings of ICML 2005*, pages 1012–1019. ACM, 2005.

[39] D. Zhang, Y. Liu, and L. Si. Serendipitous learning: learning beyond the predefined label space. In *Proceedings of the 17th ACM SIGKDD*, pages 1343–1351. ACM, 2011.