

Annotating Web Images using NOVA: NOn-conVex group spArsity

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

Ying Yuan
College of Computer Science,
Zhejiang University
Hangzhou, Zhejiang, China
tracy1108@cs.zju.edu.cn

Yong Rui
Microsoft Research Asia
No. 5, Dan Ling Street,
Haidian District Beijing, China
yongrui@microsoft.com

Shuicheng Yan
Department of Electrical and
Computer Engineering
NUS, Singapore
eleyans@nus.edu.sg

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

ABSTRACT

As image feature vector is large, selecting the right features plays a fundamental role in Web image annotation. Most existing approaches are either based on individual feature selection, which leads to local optima, or using a *convex* penalty, which leads to inconsistency. To address these difficulties, in this paper we propose a new sparsity-based approach NOVA (NOn-conVex group spArsity). To the best of our knowledge, NOVA is the first to introduce non-convex penalty for group selection in high-dimensional heterogeneous features space. Because it is a group-sparsity approach, it approximately reaches global optima. Because it uses non-convex penalty, it achieves the consistency. We demonstrate the superior performance of NOVA via three means. First, we present theoretical proof that NOVA is consistent, satisfying un-biasness, sparsity and continuity. Second, we show NOVA converges to the true underlying model by using a ground-truth-available generative-model simulation. Third, we report extensive experimental results on three diverse and widely-used data sets Kodak, MSRA-MM 2.0, and NUS-WIDE. We also compare NOVA against the state-of-the-art approaches, and report superior experimental results.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: Information Search and Retrieval; I.4.7 [Image Processing and Computer Vision]: Feature Measurement—*feature representation*

General Terms

Algorithms, Experimentation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

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

Copyright 2012 ACM 978-1-4503-1089-5/12/10 ...\$15.00.

Keywords

NOn-conVex group spArsity, Consistent, Group Feature Selection, Image Annotation

1. INTRODUCTION

The number of Web images already reached the order of 100 Billion. Out of these images, however, only a limited percent are annotated. Developing an effective way to accurately annotate the Web images automatically are therefore of great importance. While an image is worth a thousand words, it is also true that we can extract a large number, sometimes even more than a thousand, of visual features from an image. These features may include global features, e.g., color, texture and shape, and local features, e.g., SIFT, Shape Context and GLOH (Gradient Location and Orientation Histogram) [25]. The size of the concatenated feature vector can be large, but only a subset of features carry the biggest discriminating power. As a result, selecting the right features plays a fundamental role in image annotation with high-dimensional features.

Motivated by the recent advance in compressed sensing, *sparsity*-based feature selection approaches are developed in both machine learning and multimedia communities [27]. The basic idea of sparsity-based feature selection is to employ regularizers to induce sparsity during discriminative feature selection. Depending on if the feature selection is based on individual or group features and if the regularizers are convex or non-convex, the existing approaches can be categorized into four quadrants, as shown in Table 1, e.g., *convex individual-sparsity*, *convex group-sparsity*, *nonconvex individual-sparsity*, *nonconvex group-sparsity*.

- The *lasso* (least absolute shrinkage and selection operator) [24] and nonnegative garrote [4] fall into the first quadrant (i.e., *convex individual-sparsity*), which deals with individual feature selection using a convex relaxation such as the ℓ_1 -norm.
- The group *lasso* [30], sparse group *lasso* [15], composite absolute penalty (a generalization of the group *lasso*) [33], exclusive group *lasso* [34], and overlapped group *lasso* [18] are in the second quadrant (i.e., *convex group-sparsity*) which model the features as groups, but still use a convex relaxation.

- Residing in the fourth quadrant (i.e., *nonconvex individual-sparsity*) are the smoothly clipped absolute deviation (SCAD) [12], adaptive *lasso* [35], log-sum penalty [6], pseudo l_q -norm with $q < 1$ [16], and the minimax concave (MC) penalties [32]. While they utilize a non-convex penalty to obtain prediction consistency, it only works on top of individual features.
- Note that approaches in quadrant 3 (i.e., *nonconvex group-sparsity*) are still missing.

While good progress has been made over the years in various domains and applications, the above mentioned existing approaches in quadrants 1, 2 and 4 suffer from one or both of the following difficulties.

- Note that the original high-dimensional heterogeneous features are inherently partitioned into different groups, e.g., color vs. SIFT, and different groups of features describe different aspects of visual characteristics. That is, the Web image annotation problem is inherently the problem of selecting grouped variables (features) for accurate prediction in regression, a classic statistical problem as described in [30]. Different groups of features have different intrinsic discriminative power to characterize the high-level semantics for different images [7][9][28][17]. Approaches in quadrants 1 (*convex individual-sparsity*) and 4 (*nonconvex individual-sparsity*) are designed for selecting individual features, not for group features as in Web image annotation. They make selection based on the strength of individual features instead of the strength of groups of features, which result in selecting unnecessary features. Furthermore, they lead to local optimal instead of the global one – depending on how the features orthonormalized, they converge to different solutions [30].
- Because approaches based on convex regularizers, i.e., the approaches in quadrants 1 (*convex individual-sparsity*) and 2 (*convex group-sparsity*), are computationally efficient, and are very popular [24][4][30][15][18]. However, they do not lead to *consistent* prediction. That is, the correct sparse subset of the relevant features cannot be identified asymptotically with large probability [21][26].

As pointed out in [12][36], the *lasso* as well as other *lasso*-type methods are sub-optimal in model selection due to their *convex* relaxation, which is prone to inducing inconsistent estimates for the data with high-dimensional features. Because *lasso* both shrinks and selects, it often relaxes the penalty on the relevant features. Furthermore, in Wei *et al.*'s work [26], they derived a similar theoretical proof that group *lasso* in general is not selection consistent.

Given the data x with p input features, $\mathbf{x} = (x_1, \dots, x_p)^T \in \mathbb{R}^p$, the interesting question is whether there is a *true* model that can select all the necessary features from x to denote its corresponding semantic. Assume a *true* (an *oracle*) model can be found to select out the features from x and the coefficients of the selected features are denoted as $A = \{m : \beta_m^* \neq 0\}$ and $|A| = p_0 < p$. If $\hat{\beta}(\delta)$ is the estimated coefficients produced by a fitting procedure δ , the question is whether the selection result by δ is the same as that of the result by the

true model. According to [12][35], δ is called an *oracle* model to produce a consistent selection result if $\hat{\beta}(\delta)$ has the following oracle properties: δ can identify the correct subset, $\{m : \beta_j \neq 0\} = A$; and δ has the optimal estimate rate. This oracle properties imply that the penalty function must be singular at the origin and *non-convex* over $(0, \infty)$.

To address the above two key difficulties, in this paper we propose a brand new approach, filling the empty space in the third quadrant (*nonconvex group-sparsity*). We call it NOVA (NON-conVEX group spArsity). Figure 1 illustrates the algorithm flow of our proposed NOVA, including feature extraction, feature selection and image annotation. In Figure 1, we can see NOVA can select out the important groups of features for image annotation. The strength of the proposed NOVA endows the feature space with additional group structure into a non-convex regularizer for consistent group selection due to its *oracle* properties. To the best of our knowledge, NOVA is the first to introduce non-convex penalty for group selection in high-dimensional heterogeneous features space.

The rest of the paper is organized as follows. In Section 3, we present detailed algorithm of NOVA, including problem formulation and step-by-step derivation. In Section 4, we give theoretical proof that NOVA is consistent, satisfying un-biasness, sparsity and continuity. Complementing to the theoretical proof, we also report extensive experimental results in Section 5 via three diverse and widely-used data sets.

2. NON-CONVEX GROUP SPARSITY FOR FEATURE SELECTION

2.1 Notation

Suppose we are given a set of training data with n images and C labels $\{(\mathbf{x}_i, \mathbf{y}_i) \in \mathbb{R}^p \times \{-1, 1\}^C : i = 1, \dots, n\}$, where $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T \in \mathbb{R}^p$ represents a p -dimensional feature vector for the i th image, and $\mathbf{y}_i = (y_{i1}, \dots, y_{iC})^T \in \{-1, 1\}^C$ is a C -dimensional label vector. Here, we assume $y_{ij} = 1$ if the i th image has the j th label and $y_{ij} = -1$ otherwise. Since different types of features can be extracted from images, we further assume that the p features are divided into G disjoint groups. Therefore, the feature vector of the i th image \mathbf{x}_i can be rewritten as $\mathbf{x}_i = (x_{i,1}; \dots; x_{i,G})$ where $x_{i,g} \in \mathbb{R}^{d_g}$ ($g = 1, \dots, G$) is the g th feature vector of this image, and $\sum_{g=1}^G d_g = p$.

Let $X = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$ be the $n \times p$ training data matrix, and $Y = (\mathbf{y}_1, \dots, \mathbf{y}_n)^T$ the corresponding $n \times C$ label indicator matrix. In the following, we assume that the feature matrix X is centered.

2.2 Problem Formulation

For each label $c \in \{1, \dots, C\}$, we aim to estimate the vector of regression coefficients $\beta^c = (\beta_1^c; \dots; \beta_G^c)$ where $\beta_g^c \in \mathbb{R}^{d_g}$ ($g = 1, 2, \dots, G$) by minimizing the following objective function with a linear regression model:

$$l(\beta^c) = \frac{1}{2n} \sum_{i=1}^n (y_{ic} - \sum_{g=1}^G x_{i,g}^T \beta_g^c)^2 + \sum_{g=1}^G p_\lambda(\|\beta_g^c\|_2) \quad (1)$$

Table 1: Four paradigms of sparsity-based feature selection methods (i.e., *convex individual-sparsity*, *convex group-sparsity*, *nonconvex individual-sparsity*, *nonconvex group-sparsity*)

	group	individual
convex	group <i>lasso</i> [30] sparse group <i>lasso</i> [15] overlapping group <i>lasso</i> [18] exclusive group <i>lasso</i> [34] structural grouping sparsity[28]	<i>lasso</i> [24] nonnegative garrote [4]
nonconvex	NOVA	SCAD [12] adaptive <i>lasso</i> [35] log-sum penalty [6] pseudo l_q -norm with $q < 1$ [16] the minimax concave (MC) penalties [32]

where $p_\lambda(\cdot)$ is the penalty function characterized by a tuning parameter λ .

In this section, after the introduction of one of *nonconvex individual-sparsity*, e.g., smoothly clipped absolute deviation (SCAD) [12], we will describe how to introduce the group structure into SCAD and define our proposed nonconvex group-sparsity penalty, e.g., *NOVA*.

SCAD was proposed in [12] to circumvent the weakness of penalties with a convex relaxation and had interesting theoretical properties to induce a consistent selection result. SCAD is defined as follows

$$p_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j|, & \text{if } |\beta_j| \leq \lambda; \\ -\frac{(|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2)}{2(a-1)}, & \text{if } \lambda < |\beta_j| \leq a\lambda; \\ \frac{(a+1)\lambda^2}{2}, & \text{if } |\beta_j| > a\lambda; \end{cases}$$

where $a > 2$ and $\lambda > 0$ are tuning parameters.

The SCAD penalty corresponds to a quadratic spline function with two knots at λ and $a\lambda$, leaves large values of β_j not excessively penalized and makes the solution continuous [12]. Fan and Li [12] showed that the penalized likelihood estimators perform as well as the *oracle* procedure in terms of selecting the consistent features, when the regularization parameter is appropriately chosen. The significance of the *oracle* procedure is that the proposed procedures outperform the maximum likelihood estimator and perform as well as we expect. They also showed that the Bayesian risks are not very sensitive to the values of a and suggested to use $a = 3.7$, which was also used in this paper.

Traditional *nonconvex individual-sparsity* SCAD does not encode group structure during feature selection. In the real world, given one image, we can obtain p -dimensional heterogeneous features and these p -dimensional features are naturally partitioned into G disjoint groups. Similar to *convex group-sparsity* penalty group *lasso* [30], we tend to introduce the group structure into SCAD for consistent selection of group features. Here we use an $\ell_{2,1}$ -norm $\|\beta_g^c\|_2$ to substitute the ℓ_1 -norm $|\beta_j|$. The $\ell_{2,1}$ -norm encourages sparsity at the group level. In other words, if the coefficients in β_g^c are non-zero, the g th group of features are all selected out to make the c th label discernible. On the contrary, some groups of features may be dropped out if the coefficients in β_g^c are equal to zeros. Thus we can extend *nonconvex individual-sparsity* SCAD to *nonconvex group-sparsity* NO-

VA as follows

$$p_\lambda(\|\beta_g^c\|_2) = \begin{cases} \lambda\|\beta_g^c\|_2, & \text{if } \|\beta_g^c\|_2 \leq \lambda; \\ -\frac{(\|\beta_g^c\|_2^2 - 2a\lambda\|\beta_g^c\|_2 + \lambda^2)}{2(a-1)}, & \text{if } \lambda < \|\beta_g^c\|_2 \leq a\lambda; \\ \frac{(a+1)\lambda^2}{2}, & \text{if } \|\beta_g^c\|_2 > a\lambda; \end{cases}$$

for $g = 1, \dots, G$ and $c = 1, \dots, C$.

Some other *convex group-sparsity* methods such as group *lasso* [30] or group LARS [11] end up shrinking the coefficients more for the *good* variables to induce sparsity. If these selected *good* variables are strongly correlated, this effect is exacerbated, and may mistakenly include other variables to this model. On the contrary, the proposed *nonconvex group-sparsity* method NOVA can overcome these drawbacks and give rise to the consistent group selection compared with group *lasso* or group LARS.

2.3 Algorithm and Group Selection

The *nonconvex group-sparsity* penalty NOVA is singular at the origin and does not have continuous second order derivatives. In order to solve the non-convex penalized regression problem, we use the following local quadratic approximation (LQA):

$$p'_\lambda(|\beta_j|) = p'_\lambda(|\beta_j|) \text{sgn}(\beta_j) \approx \{p'_\lambda(|\beta_j|)/|\beta_j^{(0)}|\}\beta_j$$

where $\beta_j^{(0)}$ is an initial value of β_j and $\beta_j^{(0)} \neq 0$. That is to say,

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(0)}|) + 1/2\{p'_\lambda(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}(\beta_j^2 - \beta_j^{(0)2})$$

for $\beta_j \approx \beta_j^{(0)}$.

(2)

To avoid numerical instability, Fan and Li [12] suggested that if $|\beta_j|$ in (2) is very close to 0, say $|\beta_j| < \epsilon_0$ (a pre-specified value), then set $\beta_j = 0$ and delete the j th component of X in the iteration. A drawback of this approximation is the backward stepwise variable selection: if a covariate is deleted at any step in the LQA algorithm, it will necessarily be excluded from the final selected model. Furthermore, one has to choose ϵ_0 , which practically becomes an additional tuning parameter. The value of ϵ_0 potentially affects the sparsity degree of the solution as well as the speed of convergence.

In our algorithm, a similar quadratic approximation is used by substituting $|\beta_j|$ with $\|\beta_g\|_2$, $g = 1, \dots, G$. The

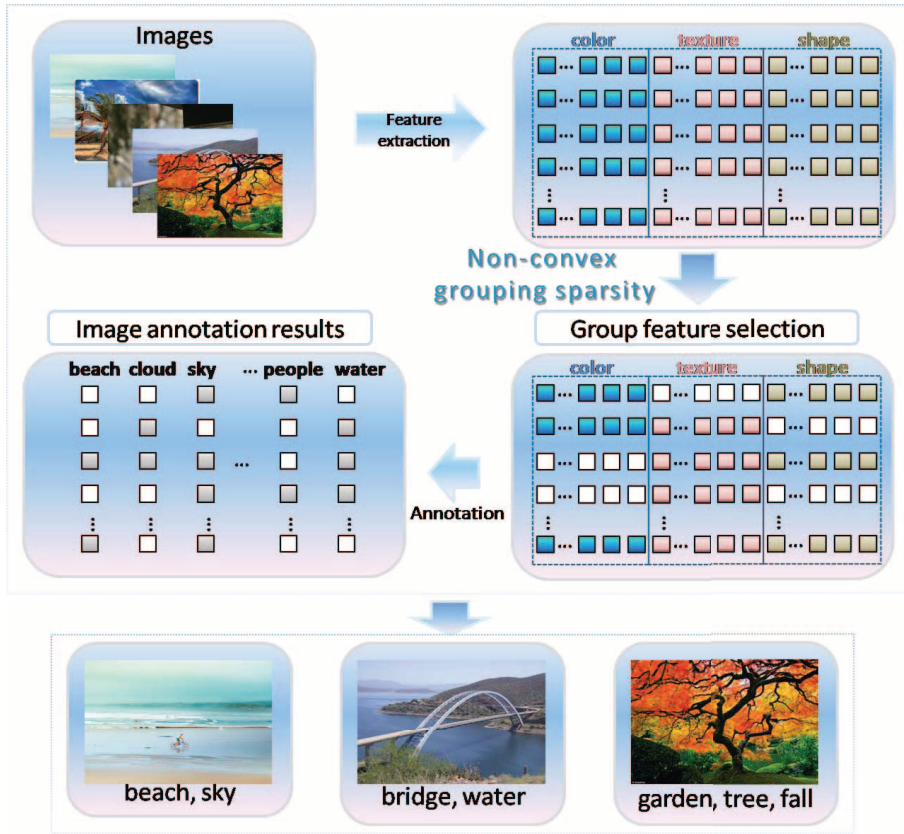


Figure 1: System overview of our proposed NOVA framework for multi-label image annotation. The NOVA can select the groups of features and train the image annotation model simultaneously.

non-convex penalty $p_\lambda(\|\beta_g\|_2)$ can be approximated as

$$p_\lambda(\|\beta_g\|_2) \approx p_\lambda(\|\beta_g^{(0)}\|_2) + 1/2\{p'_\lambda(\|\beta_g^{(0)}\|_2)/\|\beta_g^{(0)}\|_2\}(\beta_g^T \beta_g - (\beta_g^{(0)})^T \beta_g^{(0)}), \text{ for } \beta_g \approx \beta_g^{(0)},$$

and the Newton-Raphson algorithm can be conducted.

Under this approximation, the equation (1) is transformed as

$$\min_{\beta^c} \frac{1}{2n} (Y - X\beta^c)^T (Y - X\beta^c) + \frac{1}{2} (\beta^c)^T \Sigma \beta^c \quad (3)$$

where

$$\Sigma = \text{diag}\{p'_\lambda(\|\beta_1^{c(0)}\|_2)/\|\beta_1^{c(0)}\|_2 e_{d_1}^T, \dots, p'_\lambda(\|\beta_G^{c(0)}\|_2)/\|\beta_G^{c(0)}\|_2 e_{d_G}^T\} \quad (4)$$

is a $p \times p$ matrix. Each item $p'_\lambda(\|\beta_g^{c(0)}\|_2)/\|\beta_g^{c(0)}\|_2$ of Σ is repeated d_g times where d_g is the dimension of the g th feature as aforementioned.

The penalized equation (3) is a quadratic problem and can be solved by

$$(X^T X + \Sigma)\beta^c = \frac{1}{n} X^T Y \quad (5)$$

We summarize the proposed algorithm in Algorithm 1. As in the maximum likelihood estimation (MLE) setting, with the good initial value $\beta^{c(0)}$, the one-step procedure can be as efficient as the fully iterative procedure, namely, the

Algorithm 1: Group Selection with Non-convex group sparsity (NOVA)

- 1 **For** each label c ($c = 1, 2, \dots, C$) **do**
 - 2 Initialize $\beta^{c(0)}$;
 - 3 **For** $k = 1, \dots$
Obtain $\Sigma^{(k)}$ by (4) and solve $\beta^{c(k)}$ by (5) until convergence;
end
 - 4 **For** $g = 1, \dots, G$ **if** $\|\beta_g^c\|_2 < \epsilon_1$
Set $\beta_g^c = 0$.
end
-

penalized maximum likelihood estimator, when the Newton-Raphson algorithm is used [2]. Therefore, in the initialization step, we utilize a ridge regression to get an approximate initial estimation of β^c . In step 4, if some $\|\beta_g^c\|_2$ is very close to zero, that is to say, smaller than a certain threshold ϵ_1 , we set $\beta_g^c = 0$ and treat the g th group of features irrelevant with the c th label. In our algorithm, we set ϵ_1 to 10^{-3} .

2.4 Multi-label Boosting for Annotation

The introduction of correlations between multiple tags can improve the performance of multi-label annotation. As one of approaches to learn the correlations between two vari-

Algorithm 2: Multi-label Boosting by Structured Sparse Canonical Correlation Analysis (S²C&W)

- 1 Perform structured sparse canonical correlation analysis on selected features X and labels Y ;
 - 2 Output u_{x_i} and v_{y_i} ($i = 1, 2, \dots, C$);
 - 3 Compute ρ_i and s_i by (6) and (7);
 - 4 Form matrix $S = \text{diag}(s_1, \dots, s_C)$;
 - 5 Form matrix $V = (v_{y_1}, \dots, v_{y_C})^T$;
 - 6 Compute matrix $B = V^{-1}SV$;
 - 7 Compute the estimated indicators $\tilde{Y} = \hat{Y}B$.
-

ables, the curds and whey (C&W) [5] has been conducted for multi-label annotation[29] [28]. C&W builds up the connection between multiple response regression and canonical correlation analysis, and can be used to boost the performance of multi-label prediction given by the prediction results from the individual regression of each label.

Let $\beta = (\beta^1, \dots, \beta^C) \in \mathbb{R}^{p \times C}$ be the coefficient vector output by Algorithm 1, and we can get the predicted vector $\hat{Y} = X\beta$. According to [5][23], a more accurate prediction \tilde{Y} can be inferred by a linear combination $\tilde{Y} = \hat{Y}B$ after the introduction of the significant correlations between labels.

We can derive the estimates of the matrix $B \in \mathbb{R}^{C \times C}$ that take the form $B = V^{-1}SV$ where V is the $C \times C$ matrix whose rows are the label canonical coordinates output by canonical correlation analysis (CCA) and $S = \text{diag}(s_1, \dots, s_C)$ is a diagonal *shrinking* matrix which can be estimated by a generalized cross-validation (GCV) approach.

The solution by [5] of s_i is

$$s_i = \frac{\rho_i^2}{\rho_i^2 + \gamma(1 - \rho_i^2)} \quad (6)$$

with

$$\rho_i = \frac{u_{x_i}^T X^T Y v_{y_i}}{\sqrt{(u_{x_i}^T X^T X u_{x_i})(v_{y_i}^T Y^T Y v_{y_i})}} \quad (7)$$

and $\gamma = p/n$. The matrix V can be expressed as $V = (v_{y_1}, \dots, v_{y_C})^T$. Different from traditional curds and whey (C&W) method, the label canonical coordinates u_{x_i} and v_{y_i} are obtained by the the structured sparse canonical correlation analysis [8] in this paper, referred as S²C&W, since we can incorporate the rich prior structural information among label space. We summarize the multi-label annotation by S²C&W in Algorithm 2.

3. JUSTIFICATION ON CONSISTENCY

The consistent selection of groups of features is a major concern in this paper. This section provides both theoretical and experimental evidence that our NOVA does produce a consistent group selection without compromising classification accuracy or computational efficiency.

3.1 The Oracle Properties of NOVA

A good penalty that induces consistent feature selection should result in an estimator with an *oracle* property: unbiasedness, sparsity and continuity [12]. That is to say, the penalty function is bounded by a constant to produce nearly unbiased estimates for large coefficients, be singular at

the origin which is also said to be a thresholding rule of the resulting estimator to produce sparse solutions, and be continuity by certain conditions, which is also said to be stable of the model.

Now we consider the first order derivative of $p_\lambda(\|\beta_g^c\|_2)$ with respect to β_g^c , which is

$$p'_\lambda(\|\beta_g^c\|_2) = \lambda \{I(\|\beta_g^c\|_2 \leq \lambda) + \frac{(a\lambda - \|\beta_g^c\|_2)_+}{(a-1)\lambda} I(\|\beta_g^c\|_2 > \lambda)\}.$$

The *oracle* property of NON-conVex group sparsity (NOVA) can be proved as

1. The sufficient condition for unbiasedness is that when $\|\beta_g^c\|_2$ is sufficiently large, $p'_\lambda(\|\beta_g^c\|_2) = 0$, which is obvious satisfied.
2. The sufficient condition for the thresholding rule is that the minimum of the function $\|\beta_g^c\|_2 + p'_\lambda(\|\beta_g^c\|_2)$ is positive, which is satisfied with the assumption $a > 2$.
3. The sufficient and necessary condition for continuity is that the minimum of the function $\|\beta_g^c\|_2 + p'_\lambda(\|\beta_g^c\|_2)$ attained at 0. Obvious this condition can be satisfied when $\|\beta_g^c\|_2 = 0$.

The proposed NOVA in equation (1) solves a quadratic optimization problem (equation (3)), which can be solved by equation (5) with a global solution. As illustrated in references [30] and [36], the individual feature selection is often trapped into a local optimal solution rather than the global optimal one.

3.2 Complexity of NOVA

The computational complexity is crucial for the successful application of an algorithm. The complexity of group lasso is $O(p + k \ln(G))$ where p is the dimension of data, G is the number of groups, and k is the sparsity number which means the features selected out by the algorithm. From the description of Algorithm 1, we can see the main time-consuming operations are the initialization step and the solving process of equation (4) and (5) whose computational complexity is $O(np^2)$ where n is the sample size of data. So the complexity of NOVA is roughly $O(np^2)$ which is the same as SCAD.

3.3 Consistent Group Selection on Synthetic Data

In this section, we numerically compare our proposed *non-convex group-sparsity* NOVA with its counterpart *convex group-sparsity* group lasso [30] [1] in terms of consistent group selection on the synthetic data.

We sampled $X \in \mathbb{R}^{n \times p}$ from a normal distribution with zero mean vector and a covariance matrix of size $n = 200$, $p = 8$ for $G = 4$ groups of size $d_i = 2$, $i = 1, \dots, G$. Then we sampled Y from $Y = X\beta + \epsilon$ with the noise ϵ which is generated from $N(0, 1)$ and the regression coefficients $\beta = (\beta_1, \dots, \beta_G)$. For the indices of the 3rd and 4th groups, we set the corresponding entries of β_3 and β_4 to be zero and the other entries are sampled from i.i.d. $N(0, 1)$.

In Figure 2, we plot the regularization paths corresponding to the aforementioned synthetic data computed by NOVA and group lasso [30]. The left subfigure shows the values of the estimated coefficients $\hat{\beta}_i$, $i = 1, \dots, G$ with the changing parameter λ in equation (1). The right subfigure shows the values computed by group lasso [1]. The dotted straight lines are the paths of prior defined regression coefficients

β , and the other four lines are the paths of predicted coefficients. Figure 2 illustrates that the non-convex NOVA regularizer is more consistent for the selection of group of features than the convex regularizer group *lasso*.

3.4 The Selection of Image Attributes

In recent years, Human-nameable visual *attributes* are taken as middle-level features to improve the performance of image classification [14][19]. To validate the effectiveness of NOVA for image attributes selection, we conduct experiments on the well-devised Animals with Attributes (AWA) data set [19] which consists of 30475 images of 50 animals classes with 85 numeric values of visual attributes for each image. The visual attributes in AWA are manually labeled. Different from the synthetic data, if the labeled attributes are taken as predictors (features), we can psychologically and physiologically judge whether the selected features are consistent (true) or inconsistent based on the ground truth in AWA.

In order to select the attributes of different classes, we define the input matrix $X \in \mathbb{R}^{n \times a}$ and $Y \in \mathbb{R}^{n \times C}$ where n is the number of samples, $a = 85$ is the number of attributes and $C = 50$ is the number of classes. We numerically compare NOVA with Linear SVM and group *lasso* [24] for the performance of feature (attribute) selection. Since there is no explicit group information in the attributes of AWA, as a special case of group *lasso* and NOVA, we consider the 85 different attributes as 85 groups which means the group size $d_g = 1$ for all $g = 1, \dots, G$.

In this section, we utilize the precision criterion to measure the performance:

$$precision = \frac{TP}{TP + FP}$$

where TP means the true positives, and FP means the false positive.

We randomly select 1000 images for training, and the remaining for testing. This process is repeated ten times to generate ten random training/test partitions, and we report the average results with their variances in Table 2.

Figure 3 shows the selection results with the values of exemplary attributes assigned to the corresponding classes. Taking the manually labeled attributes as the ground truth, it can be shown that the selected attributes by NOVA are superior to the selected ones by group *lasso*.

4. MULTI-LABEL IMAGE ANNOTATION

This section is devoted to systematically evaluating the effectiveness of our proposed NOVA for automatic multi-label image annotation on real world data. Before presenting the experiment results, we introduce the data sets as well as the criteria used for evaluation.

4.1 Experiment Data Sets

To evaluate the performance of the proposed NOVA for image annotation, we conduct experiments on two open benchmark image data sets, i.e., MSRA-MM 2.0 [20] and NUS-WIDE [10], as well as a personal image collection, i.e., Kodak [22].

The MSRA-MM 2.0 data set aims to encourage research in multimedia information retrieval and related areas. The images and videos in the data set are collected from internet search engines. The NUS-WIDE data set is a real-world web

image data set from Flickr created by National University of Singapore. We remove the images with zero label or one label for our multi-label experiments. Images in these data sets are associated with more than one labels and the labels are used as ground truth for image annotation.

For each image in the data sets, we extract its heterogeneous features and concatenate those heterogeneous features as a vector. Each kind of homogeneous features is taken as a group. We describe the three evaluated data sets with the number of groups in features which are provided by the data set as follows:

- Kodak [22] with 3590 images and 22 labels: 657-D features are divided into 7 groups. Namely, 144-D color correlogram, 16-D co-occurrence texture features, 73-D edge direction histogram, 7-D face features, 64-D color histogram, 225-D block-wise color moments, and 128-D wavelet texture features.
- MSRA-MM 2.0 [20] with 42266 images and 100 labels: 899-D features are divided into 7 groups. Namely, 144-D color correlogram, 75-D edge direction histogram, 7-D face features, 64-D color histogram, 225-D block-wise color moments, 256-D RGB and 128-D wavelet texture features.
- NUS-WIDE [10] with 209347 images and 81 labels: 1134-D features are divided into 6 groups. Namely, 144-D color correlogram, 73-D edge direction histogram, 500-D bag of words, 64-D color histogram, 225-D block-wise color moments and 128-D wavelet texture features.

4.2 Evaluation Criteria

The area under the ROC curve, called AUC, is used to measure the performance of image annotation [13]. ROC curve is used to characterize compromise relations between TP and FP. A ROC curve is a two-dimensional depiction of classifier performance which cannot be compared with each other directly for their overlap. To compare classifiers we may want to reduce ROC performance to a single scalar value representing expected performance. A common method is to calculate the area under the ROC curve, abbreviated AUC [3]. In this work, we use MacroAUC, MicroAUC scores and precision to measure the annotation performance across multiple labels.

4.3 Experiment Setup

For each data set, we randomly select $25 \times C$ images for training where C is the number of labels, and the remaining for testing. During the sampling process, each label is guaranteed to appear in at least one image. This process is repeated ten times to generate ten random training/test partitions, and we report the average results along with their variances.

For each label c , we first perform NOVA to select groups of features on the training data in order to obtain the best discriminant groups of features for label c , and then we can obtain the predicted vector \hat{Y} by the new feature vector with some kinds of groups of features are dropped out and others are kept for distinguishing different objects. Taking the correlations and interdependent among labels into account, we utilize the $S^2C\&W$ method to improve the annotation accuracy by $\hat{Y} = \hat{Y}B$.

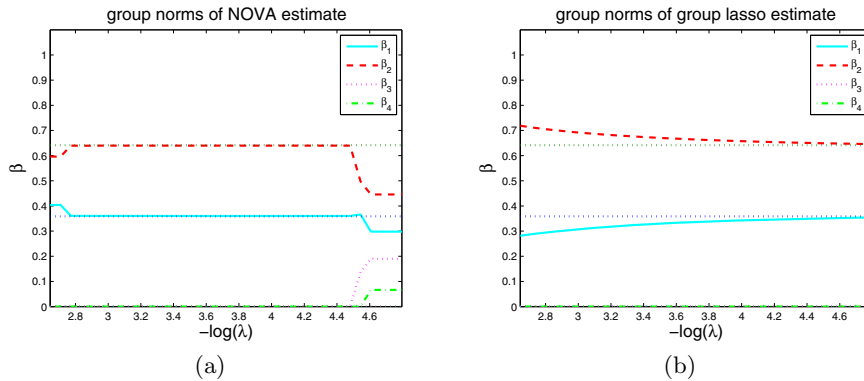


Figure 2: The regularization paths of NOVA and group *lasso*. Note that the $\|\beta_3\|$, $\|\beta_4\|$ are equal to zero, and they are overlapped on the 0-line.

Table 2: Performance of image attributes selection in terms of precision

	SVM	group <i>lasso</i>	NOVA
Precision	0.9289 ± 0.0054	0.9568 ± 0.0042	0.9784 ± 0.0156

4.4 Parameters Tuning

There are two parameters which need to be tuned for each label. The first one is λ in (1), and the second one is α in the ridge regression method for initializing $\beta^{c(0)}$. The parameters are tuned by 5-folded cross-validation based on MacroAUC for each data set.

We depict three examples of parameter tuning by the 5-fold cross validation with respect to MacroAUC in Figure 4. From this figure, we can see that the proposed algorithm is insensitive to the parameters.

4.5 Performance Comparison

As discussed before, in general, there are four paradigms of sparsity-based feature selection, namely, *convex individual-sparsity*, *convex group-sparsity*, *nonconvex individual-sparsity*, *nonconvex group-sparsity*. In order to testify the consistent selection of groups of features plays a fundamental role for image annotation, we compare our proposed NOVA with its counterpart algorithms. Without a special explanation, the annotation performance of all of compared sparsity-based are boosted by structured sparse canonical correlation analysis (S²C&W)[8]. Under such a setting, the annotation performance is solely influenced by the different schemas of sparsity-based feature selection.

The compared sparsity-based feature selection algorithms with our proposed NOVA are listed as following:

- **convex individual-sparsity:** *lasso*: Least Absolute Shrinkage and Selection Operator (*lasso*) [24] selects the important features individually and disregards the group structure in features.
- **convex group-sparsity:** **group *lasso***: group *lasso* [30] encodes the group structure in features to encourage the selection of groups of features.
- **nonconvex individual-sparsity:** **SCAD**: Smoothly Clipped Absolute Deviation (SCAD) [12] is a nonconvex penalty for individual feature selection.

All the methods are repeated ten times for ten random training/test partitions, and we report the average results

and their standard deviation. Table 3 shows the performances in terms of MacroAUC, MicroAUC and precision on three data sets. The results shown in boldface are best results.

From the results in Table 3, we can make the following observations:

- The proposed NOVA achieves the best performance of image annotation in almost all of metrics for all the three data sets thanks to its group-based sparsity and non-convex penalty.
- For MSRA-MM 2.0 data set, group *lasso* performs better than NOVA in the precision measure. But for multi-label image annotation, MacroAUC and MicroAUC are more accurate indicators of true performance. For those two measures, NOVA continues outperforming group *lasso*.
- The approaches in the first quadrant, e.g., *lasso*, have the worst performance. This is easy to understand their feature selection is performed on top of individual features, which leads to local optima, and their penalty is convex, which lead to inconsistency.
- The approaches in the second and fourth quadrants have better performance than Lasso, but worse than NOVA, as they only solved one of the two difficulties.

To visually illustrate the superior performance of NOVA, Figure 5 shows example annotation results from the MSRA-MM 2.0 data set by NOVA and group *lasso*. The annotations with underlines denote the wrong ones. We can see that group *lasso* made several mistakes, e.g., building, water, and candle. Furthermore, while not completely wrong, group *lasso* missed a few annotations from the ground truth, e.g., building and animal. A key reason for the wrong and missed annotations was because group *lasso*'s convex penalty is inconsistent as discussed in earlier sections.




Image samples with class labels	Attributes	Ground truth	group <i>lasso</i>	NOVA
giant+panda 	black:	yes	yes	yes
	white:	yes	yes	yes
	furry:	yes	<u>no</u>	yes
	longleg:	no	no	no
	claws:	yes	yes	yes
	slow:	yes	<u>no</u>	yes
	strong	yes	yes	yes
	...			
horse 	brown:	yes	yes	yes
	furry:	yes	yes	yes
	big:	yes	<u>no</u>	yes
	lean:	yes	yes	yes
	hooves:	yes	yes	yes
	tail:	yes	yes	yes
	horns	no	<u>yes</u>	no
	...			
wolf 	brown:	yes	yes	yes
	gray:	yes	yes	yes
	furry:	yes	yes	yes
	fast:	no	no	no
	insects	no	no	no
	meat:	yes	yes	yes
	smart:	yes	<u>no</u>	yes
	...			

Figure 3: Some exemplar results of attributes selection with the classes and the corresponding attributes. On the right side of the images shows the classes and the corresponding groundtruth of attributes and that selected by group *lasso* and NOVA respectively.

5. CONCLUSIONS

To address the difficulties from individual feature selection and convex penalty, this paper proposed a new sparsity-based approach termed NOVA (NON-conVex group spARsity). We have demonstrated the superior performance of NOVA via three means in the paper. First, we derived the theoretical proof that NOVA is consistent, satisfying un-biasness, sparsity and continuity. Second, we showed that NOVA converges to the true underlying model by using a ground-truth-available generative-model simulation. The comparisons between NOVA and the state-of-the-art approaches in the other three quadrants showed that NOVA achieved the best performance.

However, high-dimensional heterogeneous features extracted from the image are often embedded in a nonlinear and inseparable subspace. Some kernel-based methods such as [31] have been proposed to discern the embedded subspace but do not lead to *consistent* prediction. Therefore, non-convex kernel methods is worthy of consideration and research.

6. ACKNOWLEDGEMENTS

This work was supported in part by National Basic Research Program of China (2012CB316400), NSFC (60833006, 61070068), National Key Technology R&D Program (2011BAD24B03), National HeGaoJi Key Project (2010ZX01042-002-003).

7. REFERENCES

- [1] F. Bach. Consistency of the group lasso and multiple

kernel learning. *The Journal of Machine Learning Research*, 9:1179–1225, 2008.

- [2] P. Bickel. One-step huber estimates in the linear model. *Journal of the American Statistical Association*, pages 428–434, 1975.
- [3] A. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.
- [4] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, pages 373–384, 1995.
- [5] L. Breiman and J. Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):3–54, 1997.
- [6] E. Candes, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008.
- [7] L. Cao, J. Luo, F. Liang, and T. Huang. Heterogeneous feature machines for visual recognition. *Proceedings of the IEEE International Conference on Computer Vision*, 2009.
- [8] X. Chen, H. Liu, and J. Carbonell. Structured sparse canonical correlation analysis. 2012.
- [9] X. Chen, X. Yuan, S. Yan, J. Tang, Y. Rui, and T. Chua. Towards multi-semantic image annotation with graph regularized exclusive group lasso. In

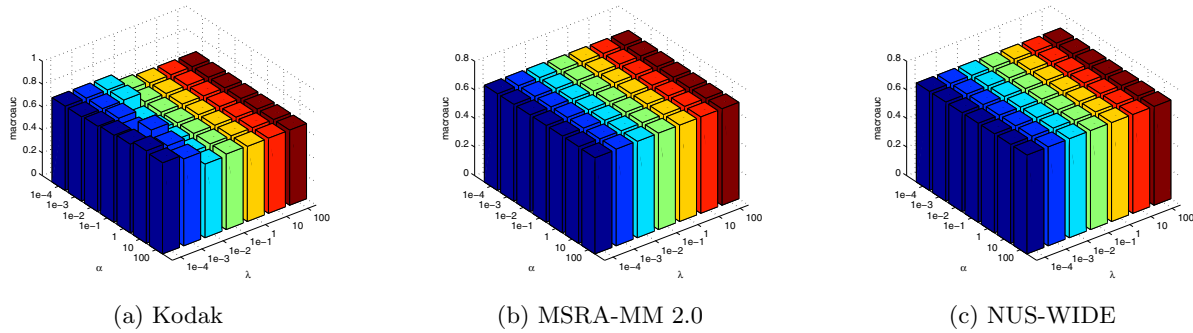


Figure 4: The change of MacroAUC scores as the regularization parameters α and λ vary in the range $[10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100]$ for the Kodak (left panel), MSRA-MM (middle panel), and NUS-WIDE (right panel) data sets.

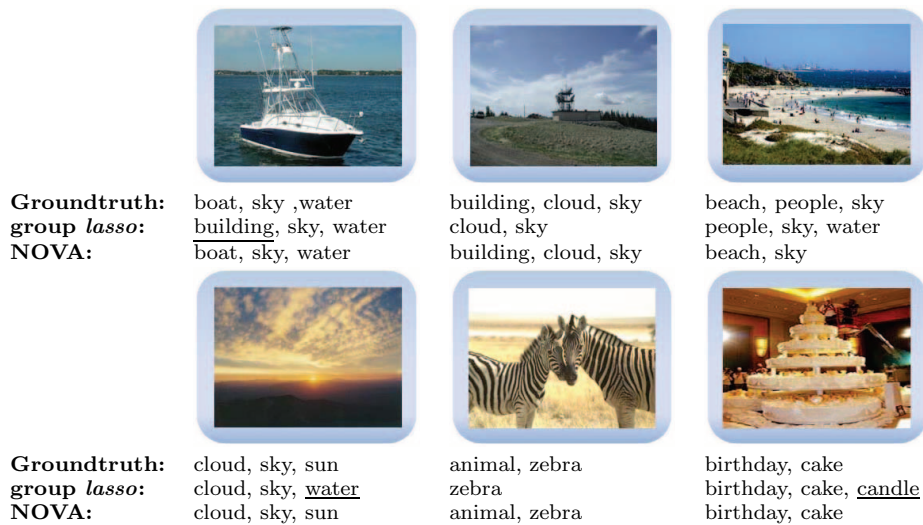


Figure 5: Image annotation results on test images from the MSRA-MM 2.0 dataset. The first row shows the corresponding groundtruths, the second row shows annotation results by *convex group-sparsity* based method *group lasso* and the third row shows the results from our *nonconvex group-based sparsity* NOVA method.

- Proceedings of the 19th ACM international conference on Multimedia*, pages 263–272, 2011.
- [10] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proceeding of the ACM International Conference on Image and Video Retrieval*, pages 1–9, 2009.
- [11] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
- [12] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.
- [13] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.
- [14] V. Ferrari and A. Zisserman. Learning visual attributes. *Proceedings of Advances in Neural Information Processing Systems*, 2008.
- [15] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. *Arxiv preprint arXiv:1001.0736*, 2010.
- [16] W. Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, pages 397–416, 1998.
- [17] Y. Han, F. Wu, Q. Tian, and Y. Zhuang. Image annotation by input-output structural grouping sparsity. *IEEE Transactions on Image Processing*, 21(6):3066–3079, 2012.
- [18] L. Jacob, G. Obozinski, and J. Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440, 2009.
- [19] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class

Table 3: Performance comparison in terms of MacroAUC, MicroAUC and Precision. The average of metrics over ten random training/test partitions are reported. The results shown in boldface are best results.

A. Multi-label annotation comparison on Kodak data set.

		MacroAUC	MicroAUC	Precision
<i>convex individual-sparsity</i>	<i>lasso</i> [24]	0.7919 ± 0.0110	0.8023 ± 0.0211	0.8751 ± 0.0201
<i>convex group-sparsity</i>	group <i>lasso</i> [30]	0.8029 ± 0.0201	0.8103 ± 0.0018	0.8813 ± 0.0128
<i>nonconvex individual-sparsity</i>	SCAD [12]	0.8231 ± 0.0089	0.8210 ± 0.0012	0.8953 ± 0.0087
<i>nonconvex group-sparsity</i>	NOVA	0.8365 ± 0.0278	0.8800 ± 0.0162	0.9131 ± 0.0317

B. Multi-label annotation comparison on MSRA-MM 2.0 data set.

		MacroAUC	MicroAUC	Precision
<i>convex individual-sparsity</i>	<i>lasso</i> [24]	0.6266 ± 0.0083	0.7826 ± 0.0121	0.8300 ± 0.0124
<i>convex group-sparsity</i>	group <i>lasso</i> [30]	0.6412 ± 0.0012	0.7906 ± 0.0010	0.9315 ± 0.0202
<i>nonconvex individual-sparsity</i>	SCAD [12]	0.6438 ± 0.0018	0.8037 ± 0.0162	0.8245 ± 0.0065
<i>nonconvex group-sparsity</i>	NOVA	0.6714 ± 0.0071	0.8255 ± 0.0035	0.8466 ± 0.0400

C. Multi-label annotation comparison on NUS-WIDE data set.

		MacroAUC	MicroAUC	Precision
<i>convex individual-sparsity</i>	<i>lasso</i> [24]	0.6540 ± 0.0112	0.7654 ± 0.0106	0.7532 ± 0.0182
<i>convex group-sparsity</i>	group <i>lasso</i> [30]	0.6612 ± 0.0040	0.7806 ± 0.1002	0.7612 ± 0.1024
<i>nonconvex individual-sparsity</i>	SCAD [12]	0.6931 ± 0.0016	0.7821 ± 0.0102	0.7631 ± 0.0338
<i>nonconvex group-sparsity</i>	NOVA	0.7029 ± 0.0045	0.7830 ± 0.0115	0.7934 ± 0.0400

attribute transfer. In *Proceedings of Computer Vision and Pattern Recognition*, pages 951–958, 2009.

- [20] H. Li, M. Wang, and X. Hua. MSRA-MM 2.0: A Large-Scale Web Multimedia Dataset. In *2009 IEEE International Conference on Data Mining Workshops*, pages 164–169, 2009.
- [21] H. Liu and J. Zhang. Estimation consistency of the group lasso and its applications. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, 2009.
- [22] A. Loui, J. Luo, S. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, and A. Yanagawa. Kodak’s consumer video benchmark data set: concept definition and annotation. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 245–254. ACM, 2007.
- [23] P. Rai and H. Daumé III. Multi-label Prediction via Sparse Infinite CCA. *Proceedings of the Conference on Neural Information Processing Systems*, 2009.
- [24] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [25] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(3):177–280, 2008.
- [26] F. Wei and J. Huang. Consistent group selection in high-dimensional linear regression. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 16(4):1369–1384, 2010.
- [27] 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*, 1(1):3–15, 2012.
- [28] F. Wu, Y. Han, Q. Tian, and Y. Zhuang. Multi-label boosting for image annotation by structural grouping sparsity. In *Proceedings of the international conference on Multimedia*, pages 15–24, 2010.
- [29] F. Wu, Y. Yuan, and Y. Zhuang. Heterogeneous feature selection by group lasso with logistic regression. In *Proceedings of the international conference on Multimedia*, pages 983–986, 2010.
- [30] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.
- [31] Y. Yuan, F. Wu, Y. Zhuang, and J. Shao. Image annotation by composite kernel learning with group structure. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1497–1500, 2011.
- [32] C. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.
- [33] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Department of Statistics, UC Berkeley, Tech. Rep*, 703, 2006.
- [34] Y. Zhou, R. Jin, and S. Hoi. Exclusive lasso for multi-task feature selection. *JMLR W&C Proceedings (AISTATS2010)*, pages 988–995, 2010.
- [35] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.
- [36] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1566, 2008.