ICIC Express Letters
Part B: Applications
Volume **7**, Number **11**, November **2016**

ICIC International ©2016 ISSN 2185-2766
pp. 2483–2490

MULTI-VIEW MIXED-NORM SPARSE CODING FOR IMAGE ANNOTATION

Miao Zang[1,2], Huimin Xu[1] and Yongmei Zhang[2]

[1]School of Information and Communication Engineering
Beijing University of Posts and Telecommunications
No. 10, Xitucheng Road, Haidian District, Beijing 100876, P. R. China
zangm@ncut.edu.cn; huimin@bupt.edu.cn

[2]School of Electronics and Information Engineering
North China University of Technology
No. 5, Jinyuanzhang Road, Shijingshan District, Beijing 100144, P. R. China
zhangym@ncut.edu.cn

Received May 2016; accepted August 2016

Abstract. *Automatic image annotation is an important research problem in computer vision. Many existing algorithms integrate multiple types of features of sample images into joint sparse coding framework to achieve better annotation performance. However, joint sparse representation only applies the sparsity to coding coefficients between the rows, which is limited in many cases, and the label information cannot be employed effectively to boost the discriminative power. In this paper, we present a multi-view mixed-norm sparse coding framework for image annotation, which integrates multiple features into sparse coding for multi-view learning problem. We introduce the mixed-norm sparsity by combining L_1-norm with $L_{\infty,1}$-norm regularization together to best represent images. In addition, the label information is considered as an additional view which leads to a simple and direct label transfer scheme. Experimental results on Corel5K and ESP Game datasets demonstrate the effectiveness of proposed method compared with the related approaches for image annotation task.*

Keywords: Multi-view learning, Image annotation, Mixed-norm regularization, Sparse coding

1. **Introduction.** Automatic image annotation refers to the task of automatically assigning relevant text labels to a given image based on its semantic content. It has become an active research topic since it is crucial for searchable databases. Image annotation is essentially a typical multi-label learning problem, in which each image could be associated with multiple labels. Many algorithms have been proposed to solve this problem by making full use of ground truth. Although great progress has been made in recent years, the problem is still challenging since an arbitrary image often captures a variety of semantic concepts, each of which would require separate detection.

Multiple features are usually employed in image annotation and classification to exploit the complementary information for performance improvement. Considerable efforts [1-6] have been made to combine information from different features. Guillaumin et al. [1] concatenate different features into a long feature vector for a KNN based image annotation. Makadia et al. [2] and Zhang et al. [3] introduce sparsity and group sparsity into feature selection to get the weight for each bin of features. Shi et al. choose more sparse and more discriminative features by exploiting the $L_{2,1/2}$-matrix norm with shared subspace learning for supervised [4] and semisupervised [5] image annotation. However, the concatenating feature cannot efficiently explore the complementary of different features because it improperly treats different features carrying different physical characteristics

[9]. Yuan et al. [6] present a multi-task joint sparse representation coding for image classification which generates different sparse representation tasks from different modalities of features and uses the constraint of joint sparsity across different tasks to enforce the robustness in coefficient representation. However, they assume that all the tasks share the same sparsity pattern, which is not applicable when the number of features increases because of the diversity of multiple features. Zhang et al. [7] propose a mixed-norm sparse representation for multi-view face recognition. It integrates multiple poses of face image by multi-view learning, and introduces a mixed-norm sparse representation coding combining typical L_1-norm and joint $L_{2,1}$-norm together to get a more flexible representation. However, image annotation is a multi-class multi-label classification problem, which cannot use their models directly. Kalayeh et al. [8] and Liu et al. [9] introduce labels as well as multiple features into multi-view learning which makes the label transfer to be very simple.

Inspired by [7-9], we propose to learn a multi-view mixed-norm sparse coding (mMSC) for image annotation. We integrate multi-view learning into the mixed-norm sparse representation framework aiming to find an optimized coefficients representation. We combine the L_1-norm with $L_{\infty,1}$-norm together and get a balance between them to find a better sparse representation for each image. Instead of using all the training samples as dictionary which leads to computation complexity and additional noise information introduced from the training samples, we introduce dictionary learning simultaneously to get a sparse linear combination of atoms from the dictionary to represent the images. Besides, we also treat the label information as an additional view, which boosts the discrimination and leads to simple label transfer scheme while not adding more computing complexity. Our experiments on Corel5K and ESP Game datasets demonstrate the effectiveness of the proposed method compared with the related methods.

The rest of this paper is organized as follows. Related work will be briefly discussed in Section 2. We will describe the details of our multi-view mixed-norm sparse coding and the label transfer scheme in Section 3. Experiments will be conducted in Section 4. We will conclude our work in Section 5.

2. **Related Work.**

Our work is closely related to the sparse coding and multi-modality learning related methods for image annotation.

2.1. **Sparse coding based image annotation.**

The past decades have witnessed the rapid development of the theory and algorithms of sparse coding and its widely and successful application in computer vision. It has also been applied to solving image annotation problems. For example, Wang et al. [10] propose to use multi-label sparse coding to automatically label images. Zhao et al. [11] explore the complementary nature of forward and backward sparse coding to find a cooperative kernel sparse representation for image annotation. Both of them use L_1-norm sparsity regularization to adaptively select training images to reconstruct test image. While these methods obtain good results, any structure prior in the observed data failed to be considered, which is often useful for image annotation. Zhang et al. [3] consider each type of feature as a group and introduce group sparsity to select features for image annotation. Gao et al. [12] treat each (sub)class of samples as a (sub)group and present a multilayer group-sparse coding to classify and annotate single-label images concurrently. They use a joint sparsity with L_2-norm inside a group and L_1-norm between the groups to encourage members of the same group to rely on the same dictionary entries. Liu et al. [9] introduce multi-view learning joint sparse coding for image annotation in which the sparsity constraint is applied between the rows of coefficient matrix to select dictionary columns sparsely. Although the joint sparsity helps to improve the annotation performance, it does not work well for the samples with large variance in the same group.

2.2. **Multi-modality learning.** In practice, it is often the case that we can obtain multiple modalities of features from the same image, such as color histogram, edge sketch and local binary patterns (LBP), characterizing different properties of an image. These different features capture different aspects. It would be beneficial if we could exploit these complementary features together for classification and annotation. Conventional approaches concatenate different features into a long vector, and adopt different feature selection methods to determine the weight for each type of feature. Recently, multi-feature learning has been introduced in many applications [13-16]. Wang et al. [13] introduce multiple features in metric learning for semantic classification and automatic tagging. Yang et al. [14] use multi-feature collaborative model for pattern classification. Yuan et al. [6] present a multi-task joint sparse representation for image classification. Sandhan and Jin [15] and Hu et al. [16] apply multi-feature joint sparse representation for gesture recognition and object tracking respectively. However, these learning frameworks do not account for the class label information, which is crucial for image annotation. Kalayeh et al. [8] and Liu et al. [9] introduce multi-view learning for image annotation. Both of them treat labels as an additional view, which exploits the discriminative information effectively and obtains better annotation performance.

3. **Proposed Method.** In this section, we will describe the formulation of our method, its optimization and the labels transfer scheme.

3.1. **Multi-view mixed-norm sparse representation model.** We propose to learn a multi-view mixed-norm sparse coding for image annotation. We employ the multiple modalities of features into the sparse coding framework for multi-view learning as well as dictionary learning. In addition, we treat the label information as an additional view to boost the discrimination power. Suppose we are given a dataset of N training sample images, each of which has K different features. Denote by $\boldsymbol{X}^{(k)} \in \mathbb{R}^{P_k \times N}$ $(k = 1, \ldots, K)$, the feature vector matrix for the kth feature from the training samples (P_k is the dimension of the kth feature), by $\boldsymbol{D}^{(k)} \in \mathbb{R}^{P_k \times N_d}$, the dictionary entries with N_d atoms for the kth feature, and by $\boldsymbol{\omega} \in \mathbb{R}^{N_d \times N}$, the shared coding coefficient matrix of training samples feature over dictionary among multiple views. Label information can be considered as another view $\boldsymbol{X}^{(K+1)} = \left[\boldsymbol{x}_1^{(K+1)}, \boldsymbol{x}_2^{(K+1)}, \ldots, \boldsymbol{x}_i^{(K+1)}, \ldots, \boldsymbol{x}_N^{(K+1)}\right] \in \mathbb{R}^{P_c \times N}$, P_c is the number of labels, $\boldsymbol{x}_i^{(K+1)} \in \mathbb{R}^{P_c}$ is the label vector of the ith image, and each entry is either 1 or 0 representing whether the occurrence of a certain label in the image or not. We aim to find a set of dictionary entries $\boldsymbol{D} = \left\{\boldsymbol{D}^{(1)}, \boldsymbol{D}^{(2)}, \ldots, \boldsymbol{D}^{(K+1)}\right\}$ and the corresponding representation coefficient $\boldsymbol{\omega}$ by solving the optimization problem:

$$\min_{\boldsymbol{\omega}, \boldsymbol{D}^{(k)}} \frac{1}{2N} \sum_{k=1}^{K+1} \|\boldsymbol{X}^{(k)} - \boldsymbol{D}^{(k)}\boldsymbol{\omega}\|_F^2 + \lambda \left[\gamma\|\boldsymbol{\omega}\|_1 + (1-\gamma)\|\boldsymbol{\omega}\|_{p,1}\right] \tag{1}$$

where $[\gamma\|\boldsymbol{\omega}\|_1 + (1-\gamma)\|\boldsymbol{\omega}\|_{p,1}]$ is the mixed-norm sparsity regularizer, and λ is the weight used to control the regularizer. γ is the tuning parameter to control the trade-off between L_1-norm regularization term and $L_{p,1}$-norm term. Generally, p can be 2 or ∞. When $\gamma = 1$, Equation (1) reduces to the typical multi-view sparse coding (mSC), and $\|\boldsymbol{\omega}\|_1 = \sum_{ij} |\omega_{ij}|$, which exploits the shared information between different views and tries to find an absolute sparse coefficient matrix to select dictionary entries based on the best representation, but it does not consider each dictionary atom (column) as a group which reflects a type of feature for a single image; when $\gamma = 0$, Equation (1) is the multi-view joint sparse coding (mJSC), and $\|\cdot\|_{p,1}$ is defined as the sum of the L_p-norm of all rows of a matrix, which encourages rows of $\boldsymbol{\omega}$ to be sparse and columns to be dense. It helps to automatically discover the row dimensionality of the weight coefficients and represent the shared information between multiple views in a single latent dimension. Although it treats the dictionary column as a whole, it requires the multiple sparse representation coefficients to share the same sparsity pattern, which omits the diversity of different features and will get less performance when the features increase. When γ is in the range $(0, 1)$, the formulation will automatically get the balance and adapt to the underlying statistics. In this paper, we use the $L_{\infty,1}$-norm regularizer since it has been proven more effective than the $L_{2,1}$-norm [17].

The optimization problem in Equation (1) is convex in $\boldsymbol{D}^{(k)}$ for a fixed $\boldsymbol{\omega}$ and vice-versa. Therefore, it can be solved by alternating between optimizing $\boldsymbol{D}^{(k)}$ with a fixed $\boldsymbol{\omega}$ and the opposite. Such process is iterated until the solutions of $\boldsymbol{\omega}$ and $\boldsymbol{D}^{(k)}$ converge to some local minimum.

By fixing $\boldsymbol{D}^{(k)}$, Equation (1) can be simplified to:

$$\min_{\boldsymbol{\omega}} \frac{1}{2N} \sum_{k=1}^{K+1} \left\| \boldsymbol{X}^{(k)} - \boldsymbol{D}^{(k)}\boldsymbol{\omega} \right\|_F^2 + \lambda[\gamma\|\boldsymbol{\omega}\|_1 + (1-\gamma)\|\boldsymbol{\omega}\|_{\infty,1}] \tag{2}$$

which is a convex function, and the first term is differentiable with Lipschitz gradient, and could be solved effectively by a variant of Nesterovs first order method which is similar to [9].

By fixing $\boldsymbol{\omega}$, Equation (1) can be simplified to:

$$\min_{\boldsymbol{D}^{(k)}} \frac{1}{2N} \sum_{k=1}^{K+1} \left\| \boldsymbol{X}^{(k)} - \boldsymbol{D}^{(k)}\boldsymbol{\omega} \right\|_F^2 \tag{3}$$

which is a convex function and can be solved by the Lagrangian method.

3.2. **Label transfer.** Since we treat labels as an additional view, the label information from the sparse code can be inferred directly. In particular, given a test image represented by multi-view features $\boldsymbol{X}_* = \left\{ \boldsymbol{x}_*^{(1)}, \boldsymbol{x}_*^{(2)}, \ldots, \boldsymbol{x}_*^{(K)} \right\}$ and the learned dictionary $\boldsymbol{D}$ from the training samples, the label view of the test image $\boldsymbol{x}_*^{(K+1)}$ can be estimated by the following two steps. First, we obtain $\boldsymbol{\omega}_*$ by solving the following convex problem:

$$\min_{\boldsymbol{\omega}_*} \frac{1}{2} \sum_{k=1}^{K} \left\| \boldsymbol{x}_*^{(k)} - \boldsymbol{D}^{(k)}\boldsymbol{\omega}_* \right\|_F^2 + \lambda[\gamma\|\boldsymbol{\omega}_*\|_1 + (1-\gamma)\|\boldsymbol{\omega}_*\|_{\infty,1}] \tag{4}$$

Then, we can get label view of the testing image by

$$\boldsymbol{x}_*^{(K+1)} = \boldsymbol{D}^{(K+1)}\boldsymbol{\omega}_* \tag{5}$$

The top 5 values of the label view can be considered as possible labels.

4. **Main Results.**

In this section, we will evaluate the proposed method for image annotation experimentally. We compare the proposed mMSC with related sparse coding algorithms including multi-view sparse coding (mSC) and multi-view joint sparse coding (mJSC) with $L_{\infty,1}$-norm as well as some related state-of-arts including multi-label sparse coding (MSC) [10], LASSO [2] and group sparse coding (GS) [3]. For MSC, we concatenate 15 different features into a long feature vector. For all the multi-view method, we use labels as an additional view. Parameter λ is tuned in the range $\{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$ corresponding to the best F_1 value, and γ in our mMSC is tuned in the range $[0, 1]$. The number of dictionary atoms is set to be 200. We automatically annotate each image with 5 labels.

Algorithm 1. Multi-view mixed-norm sparse coding for image annotation.

```
Input:
    The $k$th view matrix of training samples $\boldsymbol{X}^{(k)} \in \mathbb{R}^{P_k \times N}$, $1 \leq k \leq K+1$;
    The $k$th feature vector of test image $\boldsymbol{x}_*^{(k)} \in \mathbb{R}^{P_k}$, $1 \leq k \leq K$;
    Paramters $\lambda \geq 0$, $0 \leq \gamma \leq 1$.
Output:
    Label view of test image $\boldsymbol{x}_*^{(K+1)}$.
    Initialize $\boldsymbol{D}^{(k)}$, $(1 \leq k \leq K+1)$ and $\boldsymbol{\omega}$, e.g., with random entries;
While not convergence do
    Update $\boldsymbol{\omega}$ with fixed $\boldsymbol{D}^{(k)}$, $(1 \leq k \leq K+1)$ by Equation (2);
    Update $\boldsymbol{D}^{(k)}$, $(1 \leq k \leq K+1)$ with fixed $\boldsymbol{\omega}$ by Equation (3);
end while
    Learning sparse coefficients $\boldsymbol{\omega}_*$ for the test image with learned dictionary $\boldsymbol{D}^{(k)}$
    $(1 \leq k \leq K)$ and multiple features of test image $\boldsymbol{x}_*^{(k)}$ $(1 \leq k \leq K)$ by Equation
    (4);
    Get $\boldsymbol{x}_*^{(K+1)}$ by Equation (5) using learned dictionary $\boldsymbol{D}^{(K+1)}$ and coding coef-
    ficients $\boldsymbol{\omega}_*$.
```

4.1. **Dataset and evaluation metrics.** We perform experiments on two popular datasets: Corel5K [18] and ESP Game [19]. Corel5K contains 5,000 images including 4,500 training images and 500 testing images. Each image is manually annotated with 3-5 labels from a dictionary of 374 keywords, and 3.5 keywords on average. ESP Game contains 20,770 images including 18,689 training images and 2,081 testing images. Each image has up to 15 labels from a dictionary of 268 keywords, and 4.7 keywords on average.

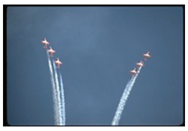
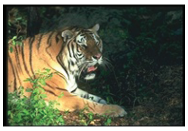
We use the publicly available features provided by [1], which consists of 15 features representing each image including a GIST feature, 2 Hue and 2 SIFT features (extracted on dense grids and Harris-Laplacian interest points respectively, represented as DenseHue, HarrisHue, DenseSIFT, and Harris-SIFT), and 6 special histogram features (computed over a 3 × 1 horizontal decomposition of the image, represented as DenseSIFTV3H1, Harris-SIFTV3H1, DenseHueV3H1, HarrisHueV3H1, RGBV3H1, LabV3H1 and HSVV3H1). Following the evaluation metrics used in [3,8], we measure the annotation performance by average precision (AP), average recall (AR) across all labels, and the number of labels with non-zero recall (N+) as well as F_1 measure.

Table 1 presents some examples of the predicted annotations produced on Corel5K and ESP Game datasets by our method. The differences between predicted and ground truth labels are marked in italic font. We notice that, in many cases, some predicted labels not contained in the ground-truth label set can still explain the image well, such as "turn" in the first image. That is because the ground-truth labels are not completed, and shows the effectiveness of our proposed method for automatic image annotation task.

4.2. **Results.** Table 2 demonstrates the performance of our proposed method compared with the related methods on both datasets. As can be seen, our method achieves the best performance on all the evaluation metrics. In particular, we find all the multi-view methods are better than MSC with concatenating long vector which exploits label information by multi-label linear embedding. This validates that the multi-view learning can exploit the image features and labels more effectively and helps to find a better representation for each image. In addition, based on the same multi-view learning framework, mixed-norm sparse coding method outperforms the other two sparse coding ones, which shows the power of mixed-norm sparse coding to select atoms dynamically.

Figure 1 demonstrates the relation between parameter γ and the annotation performance using our method. The horizontal axis represents the tuning parameter γ, and the

Table 1. Comparison of predicted labels with ground truth labels for images from Corel5K and ESP Game datasets

Images from Corel5K					
Ground truth labels	cars, formula, tracks, wall	sculpture, sphinx, statue, stone	field, foals, horses, mare	jet, plane, sky, smoke	bengal, cat, forest, tiger
Predicted labels	cars, formula, tracks, wall, *turn*	sculpture, sphinx, statue, stone, *bear*	field, foals, horses, mare, *grass*	jet, plane, sky, smoke, *prop*	bengal, cat, forest, tiger, *tree*

Images from ESP Game					
Ground truth labels	pink, girl, bookmark, read	grass, goat, animals, man, hat	fingers, type, keyboard	blue, boy, school, paper, smile	house, plant, tree, window, fence, building
Predicted labels	*blue*, girl, *book*, read, pink	grass, *dog*, man, animal, *ground*	finger, keyboard, type, *hand*, *black*	blue, boy, *girl*, paper, smile	building, tree, window, hourse, plant

Table 2. Annotation results comparision on two Datasets. MSC* refers to our implementation of [10] using our features concatenated as a long vector.

	Corel5K				ESP Game			
method	AP	AR	N+	F1	AP	AR	N+	F1
LASSO [2]	0.24	0.29	127	0.263	0.21	0.24	224	0.224
MSC*	0.27	0.32	140	0.293	0.22	0.24	220	0.23
GS [3]	0.30	0.33	146	0.314	–	–	–	–
mSC	0.28	0.33	143	0.303	0.26	0.25	230	0.255
mJSC	0.29	0.33	148	0.309	0.24	0.24	228	0.240
mMSR	0.31	0.35	155	0.329	0.27	0.26	236	0.265

vertical axis represents the value of F_1. It also verifies that mixed-norm sparse coding ($0 < \gamma < 1$) improves the annotation performance more or less compared with SC($\gamma = 1$) and JSC($\gamma = 0$). We can see that for the best value of F_1, γ selected for ESP Game dataset is larger than that for Corel5K dataset. That may be because the test images in ESP Game dataset have more variation which needs more flexible atom selection by L_1-norm.

5. **Conclusions.** This paper presents a multi-view mixed-norm sparse coding framework for image annotation problems. The main contribution of our method is introducing mixed-norm sparse coding into multi-view learning, which encodes the multiple feature

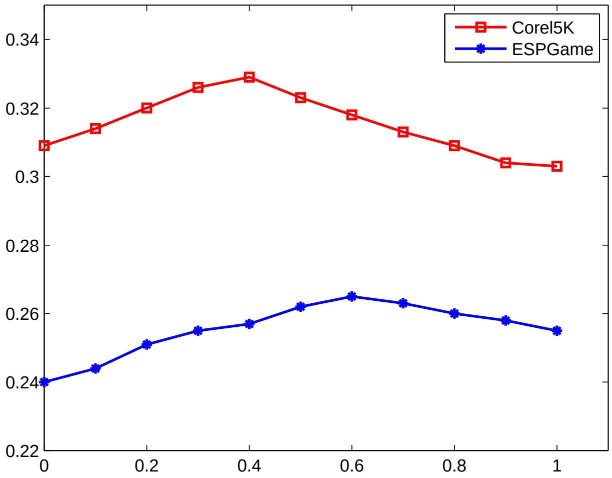


FIGURE 1. Relation between γ and F_1 performance

views of samples as well as the label view to boost the learning performance. The mixed-norm sparse coding achieves a balance between the element-wise sparsity of L_1-norm and the row-wise sparsity of $L_{\infty,1}$-norm, which helps to select dictionary atoms adaptively for better representation. Experimental results show the improvement of our method over the related methods for image annotation task. Our future research direction is to apply multi-view mixed-norm sparse coding into semi-supervised image annotation task.

Acknowledgment. This work is supported by the Chinese National Natural Science Foundation (61371143), the Beijing Natural Science Foundation (4132026) and the Scientific Research Project of Beijing Municipal Education Commission (KM201510009006). The authors also gratefully acknowledge the helpful comments and suggestions of the editors and reviewers, which have improved the presentation.

REFERENCES

[1] M. Guillaumin, T. Mensink, J. Verbeek and C. Schmid, Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation, *Proc. of the 12nd IEEE International Conference on Computer Vision*, Kyoto, Japan, pp.309-316, 2009.

[2] A. Makadia, V. Pavlovic and S. Kumar, A new baseline for image annotation, *Proc. of the 10th European Conference on Computer Vision*, Marseille, France, pp.316-329, 2008.

[3] S. Zhang, J. Huang, H. Li and D. N. Metaxas, Automatic image annotation and retrieval using group sparsity, *IEEE Trans. Syst. Man. Cybern. B*, vol.42, no.3, pp.838-849, 2012.

[4] C. Shi, Q. Ruan, S. Guo and Y. Tian, Sparse feature selection based on $L_{2,1/2}$-matrix norm for web image annotation, *Neurocomputing*, vol.151, no.1, pp.424-433, 2015.

[5] C. Shi, Q. Ruan, G. An and R. Zhao, Hessian semi-supervised sparse feature selection based on $L_{2,1/2}$-matrix norm, *IEEE Trans. Multimedia*, vol.17, no.1, pp.16-28, 2015.

[6] X. T. Yuan, X. Liu and S. Yan, Visual classification with multi-task joint sparse representation, *IEEE Trans. Image Process*, vol.21, no.10, pp.3493-3500, 2010.

[7] X. Zhang, D. S. Pham, S. Venkatesh, W. Liu and D. Phunget, Mixed-norm sparse representation for multi view face recognition, *Pattern Recognition*, vol.48, no.9, pp.2935-2946, 2015.

[8] M. M. Kalayeh, H. Idrees and M. Shah, NMF-KNN: Image annotation using weighted multi-view non-negative matrix factorization, *Proc. of the 27th IEEE Conference on Computer Vision & Pattern Recognition*, Columbus, OH, pp.184-191, 2014.

[9] W. Liu, D. Tao, J. Cheng and Y. Tang, Multiview hessian discriminative sparse coding for image annotation, *Comput. Vis. Image Und.*, vol.118, no.1, pp.50-60, 2014.

[10] C. Wang, S. Yan, L. Zhang and H. J. Zhang, Multi-label sparse coding for automatic image annotation, *Proc. of the 22nd IEEE Conference on Computer Vision & Pattern Recognition*, Miami, Florida, pp.1643-1650, 2009.

[11] Z. Q. Zhao, H. Glotin, Z. Xie, J. Gao and X. Wu, Cooperative sparse representation in two opposite directions for semi-supervised image annotation, *IEEE Trans. Image Process.*, vol.21, no.9, pp.4218-4231, 2012.

[12] S. Gao, L. T. Chia, I. W. Tsang and Z. Ren, Concurrent single-label image classification and annotation via efficient multi-layer group sparse coding, *IEEE Trans. Multimedia*, vol.16, no.3, pp.762-771, 2014.

[13] S. Wang, S. Jiang, Q. Huang and Q. Tian, Multi-feature metric learning with knowledge transfer among semantics and social tagging, *Proc. of the 25th IEEE Conference on Computer Vision & Pattern Recognition*, Providence, RI, pp.2240-2247, 2012.

[14] M. Yang, L. Zhang, D. Zhang and S. Wang, Relaxed collaborative representation for pattern classification, *Proc. of the 25th IEEE Conference on Computer Vision & Pattern Recognition*, Providence, Rhode Island, pp.2224-2231, 2012.

[15] T. Sandhan and Y. C. Jin, Frequencygrams and multi-feature joint sparse representation for action and gesture recognition, *Proc. of the 21st IEEE Conference on Image Processing*, Paris, pp.1450-1454, 2014.

[16] W. Hu, W. Li, X. Zhang amd S. Maybank, Single and multiple object tracking using a multi-feature joint sparse representation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.37, no.4, pp.816-833, 2015.

[17] A. Quattoni, X. Carreras, M. Collins and T. Darrell, An efficient projection for $L_{1,\infty}$ regularization, *Proc. of the 26th International Conference on Machine Learning*, Montreal, Canada, pp.857-864, 2009.

[18] P. Duygulu, K. Barnard, J. F. G. de Freitas and D. A. Forsyth, Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, *Proc. of the 7th European Conference on Computer Vision*, Copenhagen, Denmark, pp.97-112, 2002.

[19] L. V. Ahn and L. Dabbish, Labeling images with a computer game, *Proc. of SIGCHI Conference on Human Factors in Computing Systems*, Vienna, Austria, pp.319-326, 2004.