International Journal of Innovative Computing, Information and Control
Volume **13**, Number **4**, August **2017**

ICIC International ©2017 ISSN 1349-4198
pp. 1407–1414

MULTI-VIEW JOINT SPARSE CODING FOR IMAGE ANNOTATION

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

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

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

Received November 2016; revised March 2017

Abstract. *Automatic image annotation has become an increasingly important research topic since it is the critical and challenging component for image retrieval and management. In order to exploit the complementary information of different types of image features as well as the semantic information concurrently, we present a multi-view joint sparse coding (MVJSC) framework for image annotation, in which handcrafted features as well as deep learning based feature and label information are treated as distinct views and are exploited adaptively by multi-view learning. The adopted sparse coefficient matrix for each view is different since different view contributes differently to the final decision, and a joint sparse regularization term is introduced to ensure the similar sparse pattern across multiple views. Using the learned sparse coefficients as well as the dictionary for each view from the training images, a novel label transfer scheme is also proposed. Experiments on Corel 5K and ESP Game datasets have evaluated an improved performance of the proposed method compared with other related state-of-the-art studies.*

Keywords: Multi-view learning, Image annotation, Joint sparse coding, Dictionary learning, Deep learning

1. Introduction.

Automatic image annotation aims at automatically assigning relevant text labels to a given image reflecting its semantic content. It has become an active research topic since it has great potentials in image retrieval field. Although great progress has been made in recent years, the "semantic gap" between the image visual features and semantic labels still exists, which makes the automatic annotation problem even more challenging.

To bridge the semantic gap, considerable research efforts [1-7] have been directed to combine information from diverse features and try to exploit these complementary information for performance improvement. For example, Guillaumin et al. [1] and Gao et al. [2] concatenated different features into a long feature vector for image annotation. Makadia et al. [3] and Zhang et al. [4] introduced sparsity and group sparsity to select more sparse and discriminative features for image annotation. However, this concatenation strategy is not physically meaningful because each view has a specific statistical property [8]. Yuan and Yan [9] presented a multi-task joint sparse coding (MTJSC) for image classification which generated different sparse representation tasks from different modalities of features and exploited the constraint of joint sparsity across different tasks to enforce the robustness in coefficient representation. However, they use all the training samples as dictionary which leads to computation complexity and additional noise information introduced from the training samples. Besides, image annotation is a multi-class multi-label classification problem, which cannot use the MTJSC model directly. Liu et al. [7] introduced multi-view learning sparse coding for semi-supervised image annotation, in which each type of features as well as labels are considered as a view. However, they assume the different views share a common sparse pattern, which omit the diversity between multiple views.

Inspired by the earlier works [7,9], we propose a multi-view joint sparse coding (MVJSC) framework for image annotation. First, each feature matrix as well as the label matrix of training images is considered as a view and is sparsely coded on its associated dictionary to allow flexibility of coding coefficients. We integrate handcrafted features as well as deep learning based feature and label information into multi-view learning to obtain more robust representation. Then, a joint sparse regularization term is introduced into the multi-view learning framework to ensure similarity of each sparse pattern. Thus, we can adaptively find an optimized dictionary as well as the coefficients representation. The optimization algorithm is also proposed based on accelerated proximal gradient (APG) and K-singular value decomposition (KSVD). At last, we present our label prediction scheme based on multi-view sparse reconstruction and greedy label transfer algorithm [3]. Our experiments on Corel 5K and ESP Game datasets demonstrate the effectiveness of our proposed method and the competitive performance compared with the state-of-the-art studies.

The rest of this paper is organized as follows. Section 2 discusses the related work. Section 3 describes the details of our MVJSC, optimization and the label prediction scheme. Experimental results are reported and analyzed in Section 4. Section 5 concludes this paper.

2. Related Works.

In this section, we give a general review of sparse coding and multi-view learning related to image annotation.

2.1. Sparse coding.

In the past decades, sparse coding has been developed rapidly and is widely and successfully applied in computer vision including image annotation. Wang et al. [10] proposed to use multi-label sparse coding to automatically label images. Zhang et al. [4] introduced structural group sparsity to select features for image annotation. Gao et al. [2] presented a multilayer group-sparse coding to classify and annotate single-label images concurrently; Liu et al. [7] introduced multi-view learning sparse coding for semi-supervised image annotation. Yang et al. [11] proposed a discriminative and sparse topic model to generate latent topics such that relevant visual features and labels can be identified and irrelevant features and labels can be ignored. The success of sparse representation based classification and annotation owes to the fact that a high-dimensional image can be coded by a few representative samples in a low-dimensional manifold.

2.2. Multi-view learning.

Recently, multi-view learning has been shown to often outperform the traditional feature concatenating scheme in image annotation since it is a natural way to exploit the complementary information of multiple features and labels. Typical works include: Kalayeh et al. [6] introduced multi-view learning in nearest neighbor based image annotation, in which each type of features as well as labels are considered as a view; Liu et al. [7] integrated multi-view learning into structured sparse coding framework for semi-supervised image annotation, achieving good performance; Yang et al. [8] used single feature or the combination of several features as different views and fed each view feature into predefined deep learning model. They adopt appropriate label-specific view with the best F1-measure for annotation.

3. **Proposed Method.** In this section, we focus on multi-view joint sparse coding with dictionary learning for image annotation. Throughout this paper, given a matrix $\boldsymbol{M}$, we will use the term $\boldsymbol{M}_i$ to denote its ith column vector, and $\boldsymbol{M}_{i,\cdot}$ to denote its ith row vector.

3.1. **Multi-view joint sparse coding.** Suppose we are given a dataset of N training samples, each of which has V different features. Denote by $\boldsymbol{X}^v = [\boldsymbol{X}_1^v, \boldsymbol{X}_2^v, \ldots, \boldsymbol{X}_N^v] \in \mathbb{R}^{P_v \times N}$ $(v = 1, 2, \ldots, V)$ the feature vectors matrix for the vth feature view from the training samples (P_v is the dimension of the vth feature). The label information can be considered as another view $\boldsymbol{X}^{(V+1)} = \left[\boldsymbol{X}_1^{(V+1)}, \boldsymbol{X}_2^{(V+1)}, \ldots, \boldsymbol{X}_N^{(V+1)}\right] \in \mathbb{R}^{P_{V+1} \times N}$, P_{V+1} is the number of labels, $\boldsymbol{X}_i^{(V+1)} \in \mathbb{R}^{P_{V+1}}$ 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. Then the objective function of multi-view joint sparse coding is defined as:

$$\arg\min_{\boldsymbol{D}^v, \boldsymbol{\Omega}_i} \frac{1}{2N} \sum_{v=1}^{V+1} \sum_{i=1}^{N} \|\boldsymbol{X}_i^v - \boldsymbol{D}^v \boldsymbol{W}_i^v\|_F^2 + \gamma \|\boldsymbol{\Omega}_i\|_{1,2} \tag{1}$$

where $\boldsymbol{D}^v = \left[\boldsymbol{D}_1^v, \boldsymbol{D}_2^v, \ldots, \boldsymbol{D}_{N_d}^v\right] \in \mathbb{R}^{P_v \times N_d}$ is an overcomplete dictionary ($N_d > P_v$), and N_d is the number of dictionary atoms. $\boldsymbol{W}_i^v \in \mathbb{R}^{N_d}$ is the sparse representation coefficient of $\boldsymbol{X}_i^v$ over dictionary $\boldsymbol{D}^v$, $\boldsymbol{W}^v = [\boldsymbol{W}_1^v, \boldsymbol{W}_2^v, \ldots, \boldsymbol{W}_N^v] \in \mathbb{R}^{N_d \times N}$, and $\boldsymbol{\Omega}_i = \left[\boldsymbol{W}_i^1, \boldsymbol{W}_i^2, \ldots, \boldsymbol{W}_i^{V+1}\right] \in \mathbb{R}^{N_d \times (V+1)}$, $1 \le i \le N$. $\gamma \|\boldsymbol{\Omega}_i\|_{1,2}$ is the $L_{1,2}$ norm joint sparsity regularizer, which encourages rows of $\boldsymbol{\Omega}_i$ to be sparse. It helps us to automatically discover the dimensionality of the weight coefficients and sparsely select dictionary atoms. Furthermore, if there is shared information between several views, this regularizer will favor representing it in a single latent dimension. γ is the weight used to control the regularizer.

The optimization problem in Equation (1) can be solved by alternating between optimizing $\boldsymbol{D}^v$ with a fixed $\boldsymbol{\Omega}_i$ and the opposite. First, keeping $\boldsymbol{D}^v$ fixed, Equation (1) is simplified to:

$$\arg\min_{\boldsymbol{\Omega}_i} \frac{1}{2N} \sum_{v=1}^{V+1} \sum_{i=1}^{N} \|\boldsymbol{X}_i^v - \boldsymbol{D}^v \boldsymbol{W}_i^v\|_F^2 + \gamma \|\boldsymbol{\Omega}_i\|_{1,2} \tag{2}$$

Equation (2) can be decoupled into N distinct sub-problems, and the ith $(1 \le i \le N)$ sub-problem is formulated as follows:

$$\arg\min_{\boldsymbol{\Omega}_i} \frac{1}{2N} \sum_{v=1}^{V+1} \|\boldsymbol{X}_i^v - \boldsymbol{D}^v \boldsymbol{W}_i^v\|_F^2 + \gamma \|\boldsymbol{\Omega}_i\|_{1,2} \tag{3}$$

which can be solved by APG [12] method.

In the next step, keeping $\boldsymbol{\Omega}_i$ fixed, Equation (1) can be simplified to:

$$\arg\min_{\boldsymbol{D}^v} \frac{1}{2N} \sum_{v=1}^{V+1} \sum_{i=1}^{N} \|\boldsymbol{X}_i^v - \boldsymbol{D}^v \boldsymbol{W}_i^v\|_F^2 = \arg\min_{\boldsymbol{D}_i^v} \frac{1}{2N} \sum_{v=1}^{V+1} \|\boldsymbol{X}^v - \boldsymbol{D}^v \boldsymbol{W}^v\|_F^2 \tag{4}$$

which can be equivalent to $V + 1$ different sub-problems, and each subproblem is:

$$\arg\min_{\boldsymbol{D}^v} \frac{1}{2N} \sum_{i=1}^{N} \|\boldsymbol{X}_i^v - \boldsymbol{D}^v \boldsymbol{W}_i^v\|_F^2 = \arg\min_{\boldsymbol{D}_i^v} \frac{1}{2N} \|\boldsymbol{X}^v - \boldsymbol{D}^v \boldsymbol{W}^v\|_F^2 \tag{5}$$

This is equivalent to the dictionary update stage in dictionary learning algorithms and can be efficiently solved by the KSVD [13] method.

3.2. **Label transfer.** Since we treat labels as an additional view, the label information of the test image can be inferred without using classifiers. Specifically, given a test image represented by multi-view features $\boldsymbol{X}_* = \left\{\boldsymbol{x}_*^1, \boldsymbol{x}_*^2, \ldots, \boldsymbol{x}_*^V\right\}$ and learned dictionary $\left\{\boldsymbol{D}^1, \boldsymbol{D}^2, \ldots, \boldsymbol{D}^{V+1}\right\}$ from the training samples, the label view, i.e., the $(V+1)$th view can be estimated by the following steps. First, we obtain the sparse reconstruction coefficients $\boldsymbol{\omega}$ for each feature view by solving the following convex problem:

$$\arg\min_{\boldsymbol{\omega}^v} \frac{1}{2}\sum\nolimits_{v=1}^{V} \left\|\boldsymbol{x}_*^v - \boldsymbol{D}^v\boldsymbol{\omega}^v\right\|_2^2 + \gamma\left\|\boldsymbol{\omega}^1\ \boldsymbol{\omega}^2\ \ldots\ \boldsymbol{\omega}^V\right\|_{1,2} \tag{6}$$

Then, using each sparse coefficient matrix $\boldsymbol{\omega}^v$ $(v = 1, 2, \ldots, V)$ respectively, we can get V possible label views of the testing image by

$$\left(\boldsymbol{x}_*^{V+1}\right)_v = \boldsymbol{D}^{V+1}\boldsymbol{\omega}^v \quad (v = 1, 2, \ldots, V) \tag{7}$$

where $\left(\boldsymbol{x}_*^{V+1}\right)_v$ is the vth possible label view of the testing image. We order the values of each $\left(\boldsymbol{x}_*^{V+1}\right)_v$ and the top five values of each label view can be considered as possible labels. Then the greedy label transfer algorithm [3] is used to determine the final annotation by ordering the occurrence frequency of each label. Steps of the proposed algorithm are elaborated in Algorithm 1.

Algorithm 1. Label transfer scheme of our MVJSC

```
Input: testing image $\boldsymbol{x}_*^v \in \mathbb{R}^{P_v}$, $1 \le v \le V$; learned dictionary $\boldsymbol{D}^v$, $1 \le v \le V+1$; balancing factor $\gamma \ge 0$.
1. Learning sparse reconstruction coefficients $\boldsymbol{\omega}^v$ of test image by Equation (6).
2. For $v = 1, \ldots, V$
3.     Get possible $v$th label view of test image $\left(\boldsymbol{x}_*^{V+1}\right)_v$ by Equation (7).
4.     Order the values of $\left(\boldsymbol{x}_*^{V+1}\right)_v$ and select labels corresponding to the top five values.
5. End For
6. In the selected label set, sum the values corresponding to the same label.
7. Sort labels by the new values, and get the predicted labels with the top five values.
Output: predicted label for the test image.
```

4. **Experiments.** We present our experiments to confirm the annotation performance of the proposed method.

4.1. **Experimental settings.** We perform experiments on two popular datasets: Corel 5K [14] and ESP Game [15]. Training and testing images are selected in the manner in [1]. We use 9 publicly available visual features provided by [1] including a GIST feature, 2 Hue features (DenseHue, HarrisHue), 2 SIFT features (DenseSIFT, HarrisSIFT), and 4 histogram representations with spatial information (DenseSIFTV3H1, HarrisSIFTV3H1, DenseHueV3H1, HarrisHueV3H1). In addition, following [16], we adopt the deep learning based VGG (Visual Geometry Group) feature due to its ground-breaking results on image classification.

The two parameters γ and N_d in our MVJSC method are tuned by 5-fold cross validation on the training image set. γ is selected in the range $\{1 \times 10^e \,| e = -5, -4, -3, -2, -1, 0, 1\}$. N_d is selected in the range {4100, 4200, 4300, 4400, 4500} for Corel 5K dataset and {4500, 5000, 5500, 6000, 6500} for ESP Game. We find the best results can be reached when γ is separately set as 0.001 on Corel 5K and 0.01 on ESP Game dataset. Changing the value of N_d can increase the annotation performance slightly and generally stable. Since high dimension corresponds to expensive computing cost, N_d is set as 4300 on Corel 5K and 5000 on ESP Game dataset, respectively. Due to the random entries in initialization, we repeat all the experiments 5 times separately and report the average results.

We follow the evaluation metrics used in [4,6]. We automatically annotate each image with 5 labels, and calculate the average precision (AP), average recall (AR), F1 measure as well as the number of labels with non-zero recall (N+) for evaluation.

We make comparisons with the existing image annotation approaches related to sparse coding, including LASSO [3], MSC [10], and GS [4]. We also compare the multi-view sparse coding framework using common sparse coefficients (MVSC_CC) between the views, and the regularization term is $\gamma \|\boldsymbol{W}\|_{1,2}$. We report our method using 9 handcrafted features (MVJSC_HC), VGG feature (MVJSC_VGG) as well as both of them (MVJSC_HC+VGG).

4.2. **Experimental results.** Table 1 presents some examples of the predicted annotations produced on Corel5K and ESP Game datasets by our method. The differences in predicted labels are marked in italic font. The results in Table 1 show that, some predicted labels missed in the ground-truth label set can still reflect the image content well, such as "grass" in the first image, which shows the effectiveness of our proposed method for automatic image-annotation task.

Table 1. Comparison of predicted labels with ground truths for images from both datasets

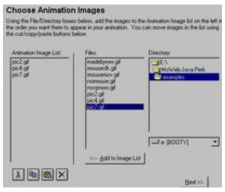
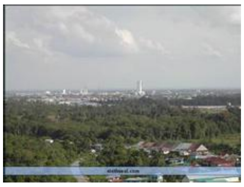
Images from Corel 5K				
Ground Truths	cars, tracks, turn, prototype	flowers, needles, blooms, cactus	sky, shore, town, windmills	sky, water, people, sand
Predicted Labels	*grass*, cars, tracks, prototype, turn	flowers, needles, blooms, cactus, *people*	sky, *water*, shore, windmills, town	water, *beach*, people, sky, sand
Images from ESP Game				
Ground Truths	animal, cat, ear, eye, gray	blue, computer, gray, screen, window	city, cloud, green, sky, tree	art, blue, building, circle, red, round, window
Predicted Labels	animal, cat, ear, eye, gray	computer, gray, *red*, screen, window	cloud, green, sky, tree, *white*	art, blue, building, circle, round

Table 2 demonstrates the performance of the proposed method compared with the state-of-the-art methods on both datasets. We can see that our MVJSC using only handcrafted features is better than or at least equal to other sparse coding based image annotation algorithms on both datasets in all the evaluation metrics. In particular, both MVSC_CC and MVJSC are clearly better than MSC, which demonstrates the effectiveness of multi-view learning and joint sparse coding; while our MVJSC is slightly better than MVSC_CC

TABLE 2. Annotation results comparison on both datasets

Method	Corel 5K				ESP Game			
	AP	AR	F1	N+	AP	AR	F1	N+
LASSO [3]	0.24	0.29	0.26	127	0.21	0.24	0.22	224
MSC [15]	0.25	0.32	0.28	136	0.35	0.23	0.28	236
GS [4]	0.30	0.33	0.31	146	–	–	–	–
MVSC_CC	0.32	0.36	0.34	152	0.36	0.28	0.32	242
MVJSC_HC	0.33	0.38	0.35	158	0.41	0.28	0.33	245
MVJSC_VGG	0.33	0.35	0.34	156	0.38	0.27	0.32	238
MVJSC_HC+VGG	0.36	0.40	0.38	170	0.44	0.30	0.36	252

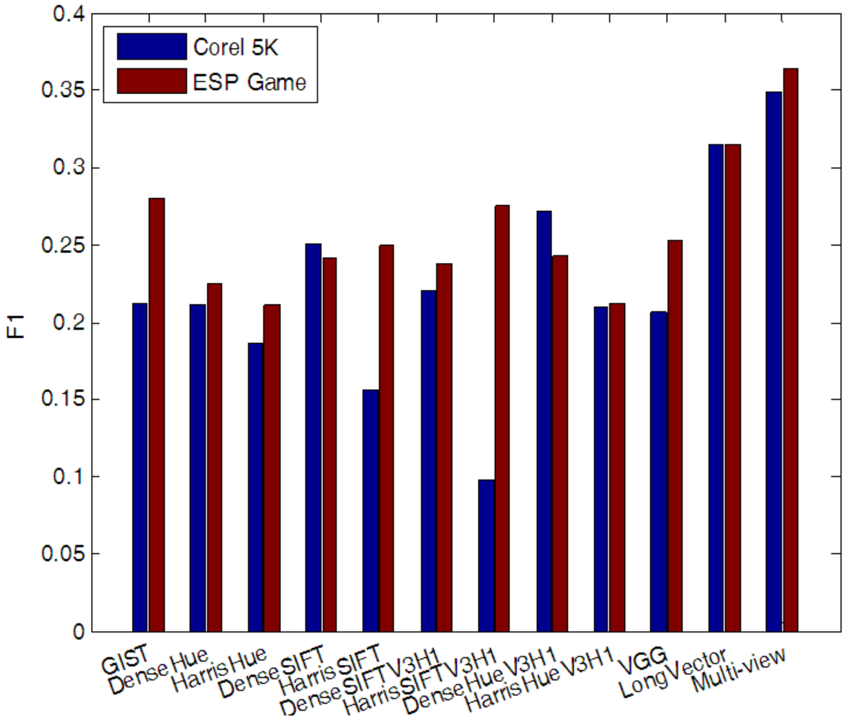


FIGURE 1. F1 comparison between single-view and multi-view methods

since the sparse coefficients in MVJSC between multiple views can be diverse and similar, which considers the fact that different views contribute differently to the pattern representation. Using handcrafted features, MVJSC_HC is slightly better than MVJSC_VGG which may because VGG feature is extracted from a pretrained neural network specific for image classification, which may not be optimal for annotation. Integrating handcrafted features and deep learning based feature, MVJSC_HC+VGG improves the performance further and get the best results, which shows multi-view learning can employ more complementary features effectively.

Figure 1 shows the performance comparison between each single-view of feature and our multi-view method. Both of them use labels as an additional view, and the label transfer schemes in both methods are the same. It can be observed that, a certain feature may achieve very different performance on different datasets. For example, HarrisSIFTV3H1 obtains nearly the best result compared with other single-view methods on ESP Game dataset, while it gets the worst performance on Corel 5K dataset. In contrast, our multiview method outperforms others including single-view and long-vector feature methods on both datasets, which demonstrates that multi-view learning can employ variant features and label information more effectively and improve annotation results significantly.

5. **Conclusions.** This paper presents a multi-view joint sparse coding framework for image annotation problems. The main contribution of our method is introducing multi-view learning into joint sparse coding, which encodes the multiple feature view of samples including handcrafted features and deep learning based feature as well as the label view to get a set of optimized dictionary along with the sparse representation for all the views. Based on such framework, we propose the label view based tag propagation method to annotate the test image. Experimental results show the effectiveness our method compared to the related state-of-the-art methods for image annotation tasks. Our future research direction is to apply kernel learning into the multi-view sparse coding framework.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (61371143), Beijing Natural Science Foundation (4132026) and 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 IEEE the 12th International Conference on Computer Vision*, Kyoto, Japan, pp.309-316, 2009.

[2] S. H. 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.

[3] A. Makadia, V. Pavlovic and S. Kumar, Baselines for image annotation, *International J. Computer Vision*, vol.90, no.1, pp.88-105, 2010.

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

[5] C. Shi, Q. Ruan, S. Guo and Y. Tian, Sparse feature selection based on L-2, L-1/2-matrix norm for web image annotation, *Neurocomputing*, vol.151, pp.424-433, 2015.

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

[7] W. Liu, D. Tao, J. Cheng and Y. Tang, Multiview Hessian discriminative sparse coding for image annotation, *Computer Vision & Image Understanding*, vol.118, pp.50-60, 2014.

[8] Y. Yang, W. Zhang and Y. Xie, Image automatic annotation via multi-view deep representation, *J. Visual Communication & Image Representation*, vol.33, pp.368-377, 2015.

[9] X. T. Yuan and S. Yan, Visual classification with multi-task joint sparse representation, *IEEE Trans. Image Processing*, vol.21, no.10, pp.4349-4360, 2012.

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

[11] L. Yang, L. P. Jing, M. K. Ng and J. Yu, A discriminative and sparse topic model for image classification and annotation, *Image & Vision Computing*, vol.51, pp.22-35, 2016.

[12] X. Chen, W. Pan, J. T. Kwok and J. G. Carbonell, Accelerated gradient method for multi-task sparse learning problem, *Proc. of IEEE the 5th International Conference on Data Mining*, Les Vegas, NV, USA, pp.746-751, 2009.

[13] M. Aharon, M. Elad and A. Bruckstein, K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation, *IEEE Trans. Signal Processing*, vol.54, no.11, pp.4311-4322, 2006.

[14] P. Duygulu, K. Barnard, J. F. G. 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.

[15] L. Von Ahn and L. Dabbish, Labeling images with a computer game, *Proc. of Sigchi Conference on Human Factors in Computing Systems*, pp.319-326, 2004.

[16] V. N. Murthy, S. Maji and R. Manmatha, Automatic image annotation using deep learning representations, *Proc. of ACM International Conference on Multimedia Retrieval*, Shanghai, China, pp.603-606, 2015.