KNN-based Image Annotation by Collectively Mining Visual and Semantic Similarities

Qian Ji[1], Liyan Zhang[2] and Zechao Li[1]
[1] School of Computer Science and Engineering, Nanjing University of Science and Technology
Nanjing, 210094, China
[e-mail: jqianxixi@163.com, zechao.li@njust.edu.cn]
[2] School of Computer Science, Nanjing University of Aeronautics and Astronautics
Nanjing, 210016, China
[e-mail: zhangliyan@nuaa.edu.cn]
*Corresponding author: Zechao Li

Received November 26, 2016; revised April 24, 2017; accepted May 28, 2017; published September 30, 2017

Abstract

The aim of image annotation is to determine labels that can accurately describe the semantic information of images. Many approaches have been proposed to automate the image annotation task while achieving good performance. However, in most cases, the semantic similarities of images are ignored. Towards this end, we propose a novel Visual-Semantic Nearest Neighbor (VS-KNN) method by collectively exploring visual and semantic similarities for image annotation. First, for each label, visual nearest neighbors of a given test image are constructed from training images associated with this label. Second, each neighboring subset is determined by mining the semantic similarity and the visual similarity. Finally, the relevance between the images and labels is determined based on maximum a posteriori estimation. Extensive experiments were conducted using three widely used image datasets. The experimental results show the effectiveness of the proposed method in comparison with state-of-the-arts methods.

Keywords: Image annotation, semantic and visual neighbors, k-nearest neighbor, semantic similarity, visual similarity

https://doi.org/10.3837/tiis.2017.09.016 **ISSN : 1976-7277**

1. Introduction

With the popular of media sharing websites and social networks, increasing numbers of people upload their images or videos to share them with their friends on the Internet, which has led to an explosive growth in the number of images. To improve the performance of visual classification and retrieval, it is necessary to assign relevant labels to images, i.e., image annotation. Image annotation is challenging due to the well-known semantic gap [1] and the expensive cost of labeled data.

Many image annotation methods have been proposed to estimate the relevance between images and labels [2-10]. These methods train annotation models using manually labeled data. They cannot address the annotation problem of a massive number of images, and so their performance is not satisfactory. As a consequence, some approaches have been proposed to train image annotation models by collecting images associated with tags from the Internet [11] [12] [13]. However, due to the imperfection of the tags, in particular incompleteness, inconsistency and error-proneness, they cannot make full use of semantic and visual information together to improve the image annotation performance.

Towards this end, we propose a novel Visual-Semantic Nearest Neighbor (VS-KNN) method for image annotation by collectively exploring visual and semantic information, as illustrated in **Fig. 1.** Specifically, given a test image, subsets of its nearest neighbors are first constructed from training images associated with each label. Due to the imperfection of image labels, another subset for each label is constructed by choosing images from the remaining images associated with the corresponding label. The visual and semantic information is jointly explored to identify the relevant images. Finally, the maximum a posteriori estimate is utilized to estimate the relevance between the test images and labels based on the selected two neighbor subsets. To evaluate the effectiveness of the proposed method for image annotation, we conducted experiments on three widely used datasets, i.e., Corel 5K [14], IAPR TC12 [15] and ESP Game [16]. The encouraging experimental results show the superiority of the proposed VS-KNN method over state-of-the-art methods.

In this paper, we propose a new method, KNN-GSR. First, we obtain subsets of the training images based on their labels. Second, we use the traditional 2PKNN method to obtain the visual neighbors of the test image in each subset. Then, we find the semantic neighbors of the visual neighbors in each subset. Finally, according to Bayes Theorem, we make use of all the neighbors to assign the importance for each label.

The key contributions of this work are as follows. 1) We propose a novel nearest neighbor method for image annotation by exploring the visual and semantic information simultaneously. 2) The imperfection of an image's labels is addressed by using its neighbors' labels to improve its labels.

The remainder of this paper is organized as follows. In Section 2, we briefly introduce related work, including discriminative models, generative models, graph-based learning methods and nearest neighbor based methods. In Section 3, we describe the proposed VS-KNN method in detail. In Section 4, we discuss the experiments and their results. Final conclusions are drawn in Section 5.

2. Related Work

Image annotation, considered a promising area in computer vision, has attracted significant attention. We can divide existing approaches into four main categories: discriminative models, generative models, graph-based learning methods and nearest neighbor methods.

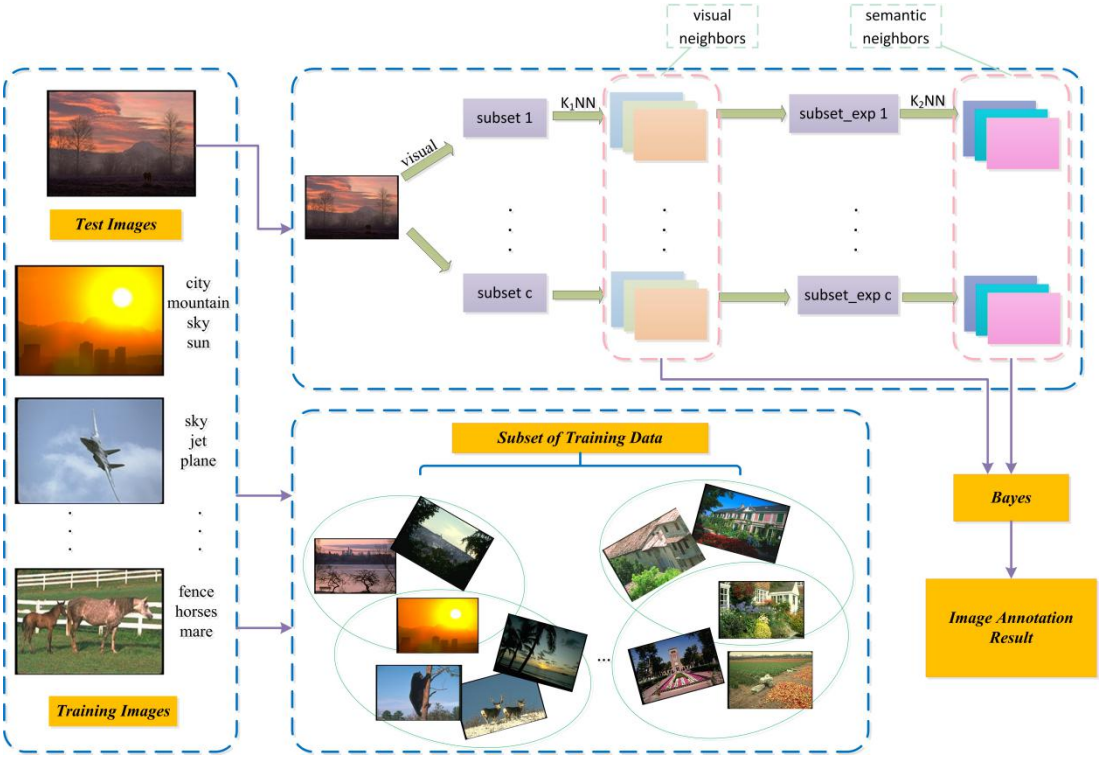


Fig. 1. The framework of our method

The discriminative models are aimed at predicting the class of an image label using several classifiers, each of which is trained for a specific label. Previously, images were usually classified into two categories [17]. However, in practice, there are more than two labels associated with an image. Therefore, it is necessary to consider image annotation as a multi-class classification problem. Carneiro et al. proposed a new probabilistic formulation for image annotation [18]. To solve the problem of confusing labels, Lavreko et al. recently presented an SVM based model by modifying the SVM hinge loss function [4]. However, such models face the problem of unbalanced training data for the positive and negative label classes, and typically the number of negative images is huge.

The generative models can be further considered as a collection of topic models and mixture models. In topic models, the images are seen as samples from a specific mixture of topics. Each topic in the mixture is a distribution of image features and corresponding annotation words. The following are instances of topic models. Barnard et al. proposed a topic model, Latent Dirichlet allocation (CorrLDA), in which the associations between image regions and labels are regarded as a mixture of latent topics [19]. Duygulu et al. presented machine translation methods, which learn a maximum likelihood association between image regions and labels in order to translate discrete image features into a vocabulary [14]. In the mixture model that is considered an important part of generative models, a joint distribution over image features and labels is defined. We can then regard the process of image annotation as learning the non-parametric density estimators over the co-occurrence of images and labels. In the continuous relevance model (CRM), each image can be divided into regions, which can be further regarded as continuous-valued feature vectors [20]. Jeon et al. proposed a cross-media relevance model (CMRM) that considers image annotation as a cross-lingual retrieval problem [21]. However, the above two models aim to maximize the generative data likelihood, which is not optimal for predictive performance.

Graph-based learning methods have been proposed to solve the image annotation issue. Pham et al. presented a new BG model, which works on a bi-relational graph of images and labels [22]. Optimal Graph Learning (OGL) [23] can address noisy label problems, and it can embed the relationships among the data points more accurately. Su et al. presented a novel graph learning based method, which propagates the labels on the graph corresponding to the K nearest neighbors of a test image [24]. However, due to their high time and space complexity, graph-based learning methods are not realistic in the real world. The performance of these models becomes even worse when the vocabulary becomes large.

Nearest neighbor methods have attracted increasing attention because of their effectiveness. We can consider the methods as sharing common labels among similar images [31]. For example, Makadia et al. proposed Joint Equal Contribution (JEC), which treats image annotation as a retrieval issue [15]. Guillaumin et al. presented a new nearest neighbor method, TagProp [25], which aims to annotate images automatically through label relevance prediction. Recently, Verma et al. proposed 2PKNN [26]. First, the method uses “image-to-label” similarities, then it uses “image-to-image” similarities. Thus, 2PKNN can make use of the two advantages mentioned above at the same time.

3. The Proposed Method

In this section, we propose a novel VS-KNN method for image annotation, which can mine visual and semantic similarities simultaneously. For visual similarity, we use the first step of 2PKNN [26] to obtain the visual neighbors of a test image. Then, for semantic similarity, we mine it using the proposed method. At the same time, due to the imperfection of labels of an image, namely incompleteness, inconsistency and error-proneness, we can also use our

method to address the problem of making the labels better. Finally, the importance of each label is assigned based on image similarity to further annotate the test image. We show a summary of the symbols used in this paper in **Table 1.**

Table 1. Descriptions of symbols

Symbols	Descriptions
X	the collection of training images
x_i	the i^{th} image in the training set
n	the number of training data
L	the dictionary of labels
l_i	the i^{th} label in the dictionary
y_i	the label set of the image x_i
I	the test image
T_i	the subset of training data that consists of all the images annotated with the label l_i
$T_{I,i}$	the K_1 nearest neighbors of the test image I in the subset T_i
T_I	the K_1 nearest neighbors of the test image I in all subset
$Tother_{I,i}$	the other images in the subset T_i except for the neighbors $T_{I,i}$ of the test image I
$S(x_j, x_k)$	the similarity between image x_j in $T_{I,i}$ and image x_k in $Tother_{I,i}$
$D(x_j, x_k)$	the visual distance between the image x_j and the image x_k
$dis(x_j, x_k)$	the tag distance between image x_j and image x_k
$P_{I,i,j}$	the K_2 images in $Tother_{I,i}$ that are the nearest to image x_j in $T_{I,i}$
β	the contribution of image x_i when using it to predict the labels
$\delta(l_k \in Y)$	whether or not the label l_k appears in the image x_i

3.1 Nearest Neighbors of a Test Image

Denote $X = \{x_1, x_2, ..., x_n\} \in R^{n \times d}$ as a collection of n images where $x_i \in R^d (1 \le i \le n)$ is the i^{th} datum. Let $L = \{l_1, l_2, ..., l_c\} \in \{0,1\}^{n \times c}$ be a dictionary consisting of c labels. The training set $T = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ contains pairs of the image x_i and its label set y_i, where each $y_i \in \{0,1\}^c$. Then, $y_i(k) = 1$ if x_i is annotated with the k^{th}

label and $y_i(k)=0$ otherwise. According to [26], we can use its method to solve the issue of class-imbalance and weak-labeling. Define $T_i \subseteq T, \forall i \in \{1,...,c\}$ as the subset of training data that consists of all the images annotated by the label $l_i (i \in [1,...,c])$. We can regard T_i as a similar group in terms of image semantics because it contains images with one common label.

Given a test image I, K_1 nearest images are selected from each semantic group by computing the visual distance between images in the group and the test image and then forming the sets $T_{I,i}$. Therefore, there are the more informative images in each $T_{I,i}$ so that the results of prediction are effectively improved. After that, we merge all of them to form the neighbors of the image I according to $T_I = \{T_{I,1} \bigcup T_{I,2} \bigcup ... \bigcup T_{I,c}\} = \bigcup_{i \in [1,...,c]} T_{I,i}$.

3.2 Improvement of Neighbors' Labels

We can never obtain perfect results, which may be the result of incomplete or noisy tagging. To address the problem, we propose a novel method that uses the other images in each subset, except for the neighbors of the test image I to improve the labels of the neighbors.

Define other images in each subset T_i, except the neighbors $T_{I,i}$ of the image I, as $Tother_{I,i} = T_i \setminus T_{I,i}$. This generates the similarity between each image in $T_{I,i}$ and in $Tother_{I,i}$ as

$$\begin{aligned} & S(x_j, x_k) = \alpha D(x_j, x_k) + (1-\alpha) dis(x_j, x_k) \\ & s.t. \quad j \in T_{I,i} \\ & \quad\;\; k \in Tother_{I,i} \end{aligned} \tag{1}$$

where $D(x_j, x_k)$ and $dis(x_j, x_k)$, respectively, denote the visual and tag distance between image x_j in $T_{I,i}$ and image x_k in $Tother_{I,i}$. We select K_2 images in $Tother_{I,i}$ that are the nearest to image x_j to form the subset $P_{I,i,j}$ and then merge them all to form a set P_I in order to further improve the neighbors' labels.

Algorithm 1. The VS-KNN method

Input:

The training image features $X \in R^{n \times d}$; The training image labels $Y \in R^{n \times c}$; The test image features $I \in R^{1 \times d}$;

Output:

The importance of each label.

1: obtain the subset of the training set $T_i \subseteq T, \forall i \in \{1,...,c\}$;

2: set t=1, T_I and P_I are both null sets;

3: repeat

compute the visual distance between each image in T_t and the test image I;

choose the K1 nearest neighbors to form the set $T_{I,t}$;

$T_I = \{T_I \cup T_{I,t}\}$;

$Tother_{I,t} = T_t \setminus T_{I,t}$;

compute the similarity between each image in $T_{I,t}$ and in $Tother_{I,t}$ using (1);

choose the K2 nearest neighbors to form the set $P_{I,t}$;

$P_I = \{P_I \cup P_{I,t}\}$;

until t>c

4: compute the posterior probability for the test image I given each label $l_k, \forall k \in \{1,...c\}$ using (2) and (3);

5: assign the weights for P_I and T_I using (4);

6: assign the importance to each label using (5) and (6);

3.3 Assign Importance to Each Label

First, for the set T_I, P_I, we respectively define the posterior probability for image I given a label l_k as

$$P_T(I \mid l_k) = \sum_{(x_i, y_i) \in T_I} \beta_{I,x_i} \cdot \delta(l_k \in Y) \tag{2}$$

$$P_P(I \mid l_k) = \sum_{(x_i, y_i) \in P_I} \exp(-S(I, x_i)) \cdot \delta(l_k \in Y) \tag{3}$$

where $\beta = \exp(-D(I, x_i))$ denotes the contribution of image x_i when we use it to predict the label l_k according to their visual similarity; and $\delta(l_k \in Y)$ denotes whether or not the label l_k appears in the image x_i such that $\delta(l_k \in Y) = 1$ if x_i is annotated with the label l_k, otherwise $\delta(l_k \in Y) = 0$.

Then, given the label l_k, we can obtain the posterior probability for I as follows:

$$P(I \mid l_k) = \theta \cdot P_T(I \mid l_k) + (1 - \theta) \cdot P_P(I \mid l_k) \tag{4}$$

According to Bayes Theorem, we can define the posterior probability for the label l_k given a test image I as

$$P(l_k \mid I) = \frac{P(l_k)P(I \mid l_k)}{P(I)} \tag{5}$$

Finally, given a test image I, the best label for it is given by

$$l^* = \arg\max_k P(l_k \mid I) \tag{6}$$

The proposed VS-KNN method is summarized **Algorithm 1**.

3.4 Computation Complexity Analysis

Now, we briefly analyze the computational complexity of our method. For a test image I, it takes $O(ck)$ to obtain T_I, and it takes $O(jL)$ to obtain P_I, where k is the number of selected visual neighbors in each subset, j is the number of selected semantic neighbors in each subset, and L is the number of corresponding labels of the semantic neighbors. The complexity of calculating the distance matrix between the n training images and m test images is $O(mn)$. Thus, the overall cost for the proposed VS-KNN is $O(m(ck + jL) + mn)$.

4. Experimental Results and Analysis

In this section, the datasets used in our experiments and the corresponding features will be described. After that, we will introduce the evaluation metrics for image annotation and provide analysis results.

4.1 Datasets Description

In the experiments, we used three widely used datasets to evaluate performance and compared the proposed method with several existing image annotation methods.

Corel 5K: This dataset has become the most common dataset used for comparing the performance of image annotation. There are 4500 training images and 499 test images. The dataset has a dictionary of 260 labels. Each image is annotated by 3.4 labels on average.

IAPR TC12: This dataset consists of 19627 images and is further split into 17665 training images and 1962 test images. The size of the dictionary is 291 with an average of 5.7 labels for each image.

ESP Game: This dataset is a collection of images annotated using an on-line game, in which the same label for an image must be assigned by two players without any communication in order to gain points. This dataset consists of 18689 training images and

2081 test images and has a dictionary of 268 labels with 4.7 labels on average.

Information on these datasets is shown in **Table 2**.

Table 2. Details for the three datasets used in this work

Dataset	No. of images	No. of labels	No. of training images	No. of test images	Labels per image (mean, maximum)
Corel 5K	4999	260	4500	499	3.4, 5
IAPR TC12	19627	291	17665	1962	5.7, 23
ESP Game	20770	268	18689	2081	4.7, 15

4.2 Features

In the experiments, we compared the performance of our method with several existing methods by using similar features with [25] and formed the features as a set. There are 15 local and global features in the set. The local features include SIFT and hue descriptors. We obtain them densely from Harris-Laplacian interest points and a multi-scale grid. The global features consist of the GIST descriptor and color histograms in RGB, LAB and HSV. It is necessary to encode the spatial information of an image by calculating the SIFT, hue descriptors and color histograms over three equal horizontal partitions for each image. We use χ^2 for the SIFT and hue descriptors, L_1 for the color histograms and L_2 for the Gist to compute the distance between two features.

4.3 Evaluation Measures

To evaluate the performance of our method, the annotation precision and recall for each label was calculated for the test set. Suppose there are m_1 images annotated with the label l_k in the ground-truth, and m_2 images predicted with the label l_k in testing which m_3 predictions are correct. Then, for the label l_k, the precision will be $\Pr = \frac{m_3}{m_2}$, and the recall will be $\mathrm{Re} = \frac{m_3}{m_1}$. We computed these values for each label and obtained the average precision P and the average recall R for each label. Then, the percentage $F1 = 2 \cdot P \cdot R/(P+R)$ by using P and R was found. In addition, N+, which denotes the number of labels that are correctly assigned to at least one test image, was also calculated. To evaluate our method, we use two criteria, $F1$ and N+.

4.4 Results and Analysis

To determine the performance of VS-KNN, we compared it with several methods, including MBRM [27], JEC [15], RF-opt [28], CCD [29], TagProp [25], FastTag [2] and 2PKNN [26]. To compare the algorithms fairly, the experiments were based on the same features and settings. We assume that each test image is annotated by 5 labels. **Table 3-5** shows the

comparison of different methods. Included are the results from [24] of the Corel 5K dataset, IAPR TC12 dataset and ESP Game dataset. "P", "R" and "F1", respectively, show the precision, recall and F1-measure. From the results, we see that the annotation accuracy of VS-KNN is superior to the state-of-the-art methods, except for 2PKNN-ML. The reasons are twofold: (1) The proposed VS-KNN method enables the mining of visual and semantic similarities simultaneously to uncover more useful information; (2) VS-KNN can address the problem of the noisy/incomplete tags by exploring the tagging information of the other images in each subset to improve the label quality.

The results in **Table 3-5** show the annotation performance of VS-KNN is lower than 2PKNN-ML. This is because 2PKNN-ML also introduces metric learning to decrease the intra-class distance and increase the inter-class distance, which also increases the computational complexity. The complexity of 2PKNN-ML is $O(m(ck+jL)+m(2\cdot I_w I_v n^4\cdot p\cdot q))$ while that of VS-KNN is $O(m(ck+jL)+mn)$, where m, n and c respectively, denote the number of test images, training images and labels. I_w and I_v are the number of iterations. We see that VS-KNN is competitive with 2PKNN-ML when comprehensively considering accuracy and time cost.

Next, we conducted experiments to study the parameter sensitivity. The results are shown in Figure 2. From the results, we see that the parameter θ in equation (4) influences the results. The proposed method achieves the best results when the value of the parameter θ is 0.6, 0.4 and 0.6 for the Corel 5K, IAPR TC12 and ESP Game datasets, respectively.

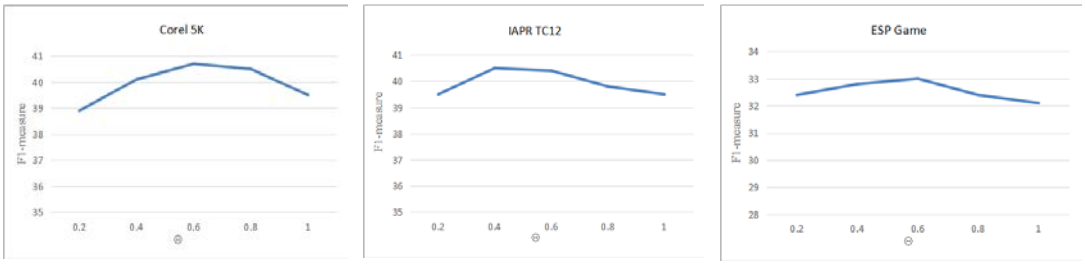


Fig. 2. The F1-score of the three datasets with different θ values

Table 3. Comparison of different methods for the Corel 5K dataset

Method	P	R	F1	N+
MBRM	24	25	24.5	122
JEC	27	32	29.3	139
RF-opt	29	40	33.6	157
CCD	36	41	38.3	159
TagProp-SD	28	35	31.1	145
TagProp-ML	33	42	37	160
FastTag	32	43	32.3	166

2PKNN	39	40	39.5	177
2PKNN-ML	44	46	45.0	191
VS-KNN	39.8	41.8	40.7	187

Table 4. Comparison of different methods for the IAPR TC12 dataset

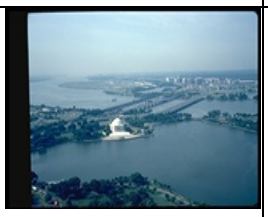
Method	P	R	F1	N+
MBRM	24	23	23.5	223
JEC	28	29	28.5	250
RF-opt	44	31	36.4	253
CCD	44	29	35	251
TagProp σSD	41	30	34.6	259
TagProp σML	46	35	39.8	266
FastTag	47	26	33.5	280
2PKNN	49	32	38.7	274
2PKNN σML	54	37	43.9	278
VS-KNN	44.6	37	40.5	278

Table 5. Comparison of different methods for the ESP Game dataset

Method	P	R	F1	N+
MBRM	18	19	18.5	209
JEC	22	25	23.4	224
RF-opt	41	26	31.8	235
CCD	36	24	28.8	232
TagProp σSD	39	24	29.7	232
TagProp σML	39	27	31.9	239
FastTag	46	22	29.8	247
2PKNN	51	23	31.7	245
2PKNN σML	53	27	35.7	252
VS-KNN	32.7	33.3	33	255

Table 6. Some examples of annotation results on the Corel 5K dataset.

image	ground truth	predicted labels	image	ground truth	predicted labels
	mountain sky sun water	sky sun water clouds people		sky jet plane	sky water clouds jet plane

	coral fish ocean reefs	water tree people coral ocean		sky tree flowers tulip	sky water tree grass flowers
	wall cars tracks formula	wall cars tracks formula water		sky water	mountain sky water clouds plane
	water bear black river	water bear snow black birds		water grass cat tiger	water tree grass cat tiger
	field horses mare foals	field horses mare foals grass		tree plane herd zebra	tree plane herd zebra grass

To illustrate the annotation results of the proposed method, some examples from the Corel 5K dataset are selected. The annotation results are shown in **Table 6**, along with the ground truth. After each image, the first column is the ground truth and the second column shows the predicted labels. Note that in the experiment, the number of predicted labels is set to five for each image. However, an image usually has fewer labels for the ground truth in the Corel 5K dataset. This shows that although the ground truth labels may be incomplete, our method can yield relevant labels. For example, in the 7th image, the predicted label ‘snow’ is relevant to the image; however, the label does not appear in the ground truth.

5. Conclusion

In this paper, we propose a novel nearest neighbor method for image annotation, VS-KNN, which is able to mine visual and semantic similarities simultaneously. For visual similarity, we obtain the visual neighbors with a first pass of 2PKNN. For the semantic similarity, our proposed method, VS-KNN, achieves its goal. At the same time, our method can improve the labeling of an image by using the labels of its visual and semantic neighbors to improve its labels. We conducted extensive experiments with three datasets. Compared to the other methods mentioned above, the proposed method shows better performance. Based on the results, we conclude that our method outperforms current state-of-the-art methods.

Acknowledgements

This work was partially supported by the 973 Program (Project No. 2014CB347600), the National Natural Science Foundation of China (Grant No. 61402228 and 61572252), and the Natural Science Foundation of Jiangsu Province (Grant No. BK20140058 and BK20150755).

References

[1] R. Bahmanyar, M.M.D.O Ambar and M. Datcu, "The semantic gap: an exploration of user and computer perspectives in earth observation images," *IEEE Geoscience & Remote Sensing Letters*, vol. 12, no. 10, pp. 2046-2050, 2015. Article (CrossRef Link)

[2] M. Chen, A. Zheng and K. Weinberger, "Fast image tagging," in *Proc. of ICML*, pp. 1274-1282, 2013. Article(CrossRef Link)

[3] H. Fu, Q. Zhang and G. Qiu, "Random forest for image annotation," in *Proc. of ECCV*, pp. 86-99, 2012. Article (CrossRef Link)

[4] Y. Verma and C. Jawahar, "Exploring SVM for image annotation in presence of confusing labels," in *Proc. of BMVC*, 2013. Article(CrossRef Link)

[5] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li and D. N. Metaxas, "Automatic image annotation using group sparsity," in *Proc. of CVPR*, pp. 3312-3319, 2010. Article(CrossRef Link)

[6] M. M. Kalayeh, H. Idrees and M. Shah, "Nmf-knn: Image annotation using weighted multi-view non-negative matrix factorization," in *Proc. of CVPR*, pp. 184-191, 2014. Article(CrossRef Link)

[7] W. Liu, D. Tao, "Multiview Hessian regularization for image annotation," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2676-2687, 2013. Article(CrossRef Link)

[8] Y. Yang, F. Wu, F. Nie, et al., "Web and personal image annotation by mining label correlation with relaxed visual graph embedding," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1339-1351, 2012. Article(CrossRef Link)

[9] R. Hong, M. Wang, Y. Gao, et al., "Image annotation by multiple-instance learning with discriminative feature mapping and selection," *IEEE Transaction on Cybernetic*, vol. 44, no. 5, pp. 669-680, 2014. Article(CrossRef Link)

[10] M. Alkaoud, I, Ashshohail, M. M. B. Ismail, "Automatic Image Annotation Using Fuzzy Cross-Media Relevance Models," *International Journal of Image and Graphics*, vol. 2, no. 1, pp. 59-63, 2014. Article(CrossRef Link)

[11] J. Tang, S. Yan, R. Hong, G. Qi and T. Chua, "Inferring semantic concepts from community-contributed images and noisy tags," in *Proc. of ACM Multimedia (MM)*, pp. 223-232, 2009. Article(CrossRef Link)

[12] J. Tang, R. Hong, S. Yan, T. Chua, G. Qi and Ramesh Jain, "Image annotation by kNN-Sparse graph-based label propagation over Noisily-Tagged Web Images," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 2, pp. 135-136, 2011. Article(CrossRef Link)

[13] Z. Li, J. Tang, "Weakly Supervised Deep Matrix Factorization for Social Image Understanding," *IEEE Trans. Image Processing*, vol. 26, no. 1, pp. 276-288, 2017. Article(CrossRef Link)

[14] P. Duygulu, K. Barnard J.F.G.D Freitas et al, "Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary," in *Proc. of CVPR*, pp. 97-112, 2002. Article(CrossRef Link)

[15] A. Makadia, V. Pavlovic and S. Kumar, "A new baseline for image annotation," in *Proc. of ECCV*, pp. 316-329, 2008. Article(CrossRef Link)

[16] L. Von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. of SIGCHI Conference on Human Factors in Computing Systems*, pp. 319-326, 2004. Article(CrossRef Link)

[17] M. Szummer and R. Picard, "Indoor-outdoor image classification," in *Proc. of IEEE international workshop on Contentbased Access of Image and Video Database*, pp. 42-51, 1998. Article(CrossRef Link)
[18] G. Carneiro, A.B. Chan, P.J. Moreno and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 3, pp. 394-410, 2007. Article(CrossRef Link)
[19] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei and M. I. Jordan, "Matching words and pictures," *Journal of machine learning research*, vol. 3, no. 2, pp. 1107-1135, 2003. Article(CrossRef Link)
[20] A. Vailaya, A. Jain and H. Zhang, "On image classification: city vs. Landscape," *Pattern Recognition*, pp. 3-8, 1998. Article(CrossRef Link)
[21] J. Jeon, V. Lavrenko and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 119-126, 2003. Article(CrossRef Link)
[22] H.D. Pham, K.H. Kim and S. Choi, "Semi-supervised Learning on Bi-relational Graph for Image Annotation," in *Proc. of ICPR*, pp. 2465-2470, 2014. Article(CrossRef Link)
[23] L. Gao, J. Song, F. Nie, et al, "Optimal graph learning with partial tags and multiple features for image and video annotation," in *Proc. of CVPR*, pp. 4371-4379, 2015. Article(CrossRef Link)
[24] F. Su and L. Xue, "Graph learning on k nearest neighbors for automatic image annotation," in *Proc. of the 5th ACM on International Conference on Multimedia Retrieval*, *ACM*, pp. 403-410, 2015. Article(CrossRef Link)
[25] M. Guillaumin, T. Mensink, J. Verbeek and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. of ICCV*, pp. 309-316, 2009. Article(CrossRef Link)
[26] Y. Verma and C. Jawahar, "Image annotation using metric learning in semantic neighborhoods," in *Proc. of ECCV*, pp. 836-849, 2012. Article(CrossRef Link)
[27] S. Feng, R. Manmatham and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *Proc. of CVPR*, pp. 1003-1009, 2004. Article(CrossRef Link)
[28] H. Fu, Q. Zhang and G. Qiu, "Random forest for image annotation," in *Proc. of CVPR*, pp. 86-99, 2012. Article(CrossRef Link)
[29] H. Nakayama, "Linear distance metric learning for large-scale generic image recognition," *PhD thesis*, The University of Tokyo, 2011. Article(CrossRef Link)
[30] S. Moran and V. Lavrenko, "Sparse kernel learning for image annotation," in *Proc. of international conference on multimedia retrieval*, pp. 113-120, 2014. Article(CrossRef Link)
[31] Z. Li, J. Tang, "Weakly Supervised Deep Metric Learning for Community-Contributed Image Retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1989-1999, 2015. Article(CrossRef Link)

Qian Ji received the BS degree in School of Computer Science and Engineering from Nanjing University of Science and Technology in 2015. She is currently a candidate for a Ph.D. degree at Nanjing University of Science and Technology. Her research interests include computer vision and information retrieval.

Liyan Zhang received the Ph.D. degree in computer science from the University of California, Irvine, in 2014. She is currently an Associate Professor with the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. Her research interests include multimedia analysis, and computer vision. She has received the Best Paper Award in ICMR 2013 and the Best Student Paper Award in MMM 2016.

Zechao Li is currently an Associate Professor at Nanjing University of Science and Technology. He received the Ph.D degree from National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences in 2013, and the B.E. degree from University of Science and Technology of China in 2008. His research interests include big media analysis, computer vision, etc. He received the Young Talent Program of China Association for Science and Technology, the Excellent Doctoral Dissertation of Chinese Academy of Sciences, the Excellent Doctoral Theses of China Computer Federation and the President Scholarship of Chinese Academy of Science.