LWTP: An Improved Automatic Image Annotation Method Based on Image Segmentation

Jianwei Niu(✉), Shijie Li, Shasha Mo, and Jun Ma

State Key Laboratory of Virtual Reality Technology and Systems,
School of Computer Science and Engineering, Beihang University,
Beijing 100191, China
niujianwei@buaa.edu.cn

Abstract. Automatic image annotation is a technique that can be used to quickly generate tags for a massive dataset based on the content of the images. Nearest-neighbor-based methods such as TagProp are successful methods which have been used for image annotation. However, these methods focus more on weights based on the distances between the images and their neighbors, and ignore the weights of the different labels which can co-occur in the same image. In this paper, an improved method is proposed for automatic semantic annotation of images, which tags rare labels more effectively by processing the label matrix of the training set. In addition, image segmentation and data-driven methods are adopted to provide differential weights to the tags in one image, to improve the accuracy of the predicted tags. Experimental results show that the proposed method outperforms many classical baseline methods and can generate better annotation results than state-of-the-art nearest-neighbor based methods.

Keywords: Image annotation · Tag propagation
Image segmentation · Local weights

1 Introduction

Automatic-annotation of images, a technology used to quickly generate information labels, has recently gained a lot of popularity in the domain of image processing. The annotation process aims to mainly elucidate the semantic information of the content of the image. The nearest-neighbor-based image annotation methods have shown to produce good results recently, and Tag Propagation (TagProp) [1] is one of them. The key component of the nearest-neighbor-based methods is transferring tags. The accuracy of this kind of methods heavily relies on the quality of the neighboring images. In TagProp, the authors used the Corel 5K [2] image dataset for the training and evaluation of the method. However, researchers [3] have noted that tags provided by users are imprecise. Besides, TagProp considers the weights of the labels only based on the distances between

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2018
I. Romdhani et al. (Eds.): CollaborateCom 2017, LNICST 252, pp. 53–62, 2018.
https://doi.org/10.1007/978-3-030-00916-8_6

the images. In reality, the weights of the labels can even vary within an image based on the image content.

To address the problems above, an improved image tagging method called the Local-Weighted Tag Propagation (LWTP) method is proposed in this paper. The proposed method, which is an extension of the TagProp method, combines the nearest-neighbor-based method and information retrieval for tag prediction. Frequency-tuned significant area detection and co-segmentation techniques are utilized to assign ranked weights to the tags of each image to improve the accuracy during the propagation process. Experimental results obtained by the LWTP outperform many baseline systems and reach the state-of-art systems, which shows the effectiveness of our method.

Specifically, the proposed method consists of the following steps: Firstly, search and crawling tools are used to supplement the image tags optionally. Secondly, the label dataset is preprocessed. Thirdly, a metric learning method is used to determine the weights of the labels based on their distances from the neighboring images. Fourthly, frequency-tuned significant area detection and co-segmentation techniques are utilized to assign ranking weights to the tags within one image. Finally, the labels of the target image are predicted based on transfer mechanism of nearest neighbor images.

The remainder of this paper is organized as follows. In the next section, we briefly review the related work. Section 3 describes the details of our approach including the steps and formulas used. Experimental results are reported in Sect. 4. Finally, our conclusion and future work are given in Sect. 5.

2 Related Work

In recent years, several learning-based methods have been applied for image annotation. The most representative methods can be broadly grouped into three categories: generative methods, discriminative methods and nearest-neighbor-based methods [4]. Some influential generative methods are the Cross Media Relevance Method (CMRM) [5], the Continuous Relevance Method (CRM) [6]and the Multiple Bernoulli Relevance Method (MBRM) [7]. These methods generally work by computing the joint probability of the tags and the visual features. In contrast, discriminative methods such as [8,9] treat image annotation as a classification problem and learn a separate classifier for each label. Nearest-neighbor-based methods treat the automatic annotation as the task of propagating labels from the neighboring images, including Joint Equal Contribution (JEC) method [10] and TagProp. More recently, new learning-based methods based on deep learning have been developed. These methods use neutral networks for feature extraction and have been discussed in [11–13]. While deep-neural-network based methods are promising, they need a large amount of aligned corpus data to train the models, which may be difficult to obtain.

Additionally, model-free image annotation methods based on information retrieval [14,15] techniques provide an inspiring solution. They use novel automatic labeling algorithms for identifying the semantic content of images by combining search and data mining techniques.

3 The LWTP Method

In this section, the proposed LWTP method for image annotation is described. Here by "Local-Weighted" it means that we consider the different label weights within a single image based on the content of the image. The LWTP method can be broken down into two main steps: training the preprocessing label matrix and predicting the image tags based on the transfer mechanism. Figure 1 shows the framework of LWTP.

LWTP is a nearest-neighbor method based on TagProp. Under the assumption that visual neighbors can transfer labels to each other, the probability density function for the presence of a label w in image i is defined as equation:

$$p(y_{iw} = +1) = \sum_{j=1}^{K} \pi_{ij} p(y_{iw} = +1|j) \tag{1}$$

where K is the number of the nearest neighbors. $p(y_{iw} = +1|j)$ indicates the presence of the label w in the image j. π_{ij} is the weight between the images i and j. Generally, the weight π_{ij} is defined as:

$$\pi_{ij} = \frac{exp(-d_\theta(i,j))}{\sum_{j'} exp(-d_\theta(i,j'))} \tag{2}$$

where $d_\theta(i,j))$ is a distance-based parameter. $d_\theta(i,j))$ is learned using metric learning. It is clear that the larger the distance is, the smaller will be the contribution of the weight of the neighbor. Since the accuracy of the method strongly depends on quality of the label dataset, it is necessary to first preprocess the label matrix.

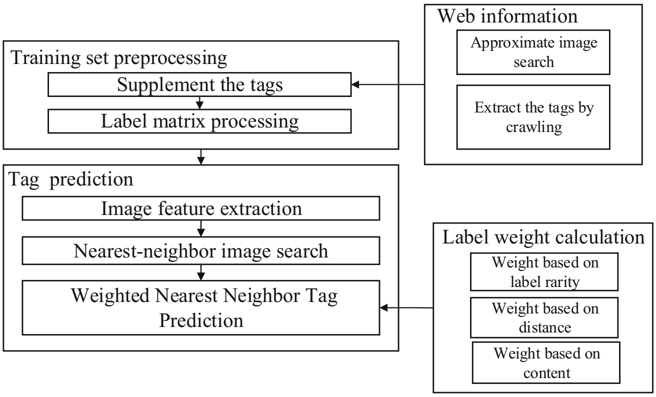

Fig. 1. Framework of the LWTP method.

3.1 Training Set Preprocessing

To tackle the deficiency in the labels, a similar image search engine[1] as well as a crawler is used to supplement the labels of the images. Then a method to balance the tag matrix is designed as following.

Let $\mathbf{L} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$ donate the training set and $\mathbf{C} = \{c_1, c_2, \ldots, c_q\}$ donate the label set, where x_i stands for the visual feature vector of each image and y_i stands for the image tag vector. $y_{iw} \in \{-1,\ +1\}$ denotes the absence/presence of a keyword c_w for an image i, to encode the image annotations. The correlation $(R_{i,j})$ between the labels i and j, which is used later, is defined as follows:

$$R_{i,j} = \frac{Col_{i,j}}{o_i + o_j - Col_{i,j}} \tag{3}$$

where o_i represents the frequency of the label c_i in the training set and $Col_{i,j}$ indicates the co-occurrence of the tags c_i and c_j. Here the label co-occurrence $Col_{i,j}$ refers to the situation where the labels c_i and c_j are both present at the same time in one image. The error function is defined as:

$$E(Y) = E_1(Y) + \alpha E_2(Y) + \beta E_3(Y) \tag{4}$$

where Y is the desired target label matrix to be processed. α and β are nonnegative integers to be solved. If the feature vectors of two images are similar, the probability of having the same labels in these two images is relatively large. So we define $E_1(Y) = \left\|YY^T - XX^T\right\|\ ^2$ where Y is the tag matrix and X is the feature vector matrix. Besides, $E_2(Y) = \left\|Y^TY - R\right\|^2$ is defined because of the assumption that there should be an association between the labels that appear together, where R stands for the correlation matrix. For example, if an image is tagged with the tags "ice", "snow", and "bear", then the probability of the label "polar" will be rather great. In addition, in order to ensure the stability of the label dataset and reduce the offset of the processed label dataset from the original dataset, $E_3(Y)$ is defined as $E_3(Y) = \|Y - Y_0\|^2$ where Y_0 donates the original label matrix. Finally, the target of the optimization is shown as:

$$\arg\min_{Y} \{\left\|YY^T - XX^T\right\|^2 + \alpha\left\|Y^TY - R\right\|^2 + \beta\|Y - Y_0\|^2\}. \tag{5}$$

The optimal solution can be found by using the gradient descent algorithm.

3.2 Tag Prediction

The final goal is to predict the annotation tags for an image by propagating the annotations of its nearest neighbors. As mentioned before, $y_{iw} \in \{-1,\ +1\}$ is used to denote the absence/presence of a keyword w in the tags of an image i.

[1] http://image.baidu.com/?fr=shitu.

The prediction of the presence of the tag w in the image i is defined as Eq. (1). The calculation of $p(y_{iw} = +1|j)$ is given below:

$$p(y_{iw} = +1|j) = \begin{cases} 1 - \varepsilon & for\ y_{jw} = 1 \\ \varepsilon & for\ otherwise \end{cases} \tag{6}$$

To avoid a prediction probability of zero, a very small factor $\varepsilon = 10^{-5}$ is used to replace any occurrence of zero. Then the objective function of the method becomes to maximize the equation:

$$L = \sum_{i,w} c_{iw} \ln p(y_{iw}) \tag{7}$$

where c_{iw} is the cost for the imbalance between the presence of the keyword and its absence. This is defined as:

$$c_{iw} = \begin{cases} \frac{1}{n^+}\ if\ y_{iw} = +1 \\ \frac{1}{n^-}\ if\ y_{iw} = -1 \end{cases} \tag{8}$$

where n^+ is the total number of the positive labels and n^- is the total number of the negative labels. As shown in Eq. (2), an approach similar to the one in [1] is used to define the weight between the image i and the image j by using the distance between them. In the whole training process, distance learning method is used to estimate the parameters. And the final logistic method uses weighted neighbor predictions as:

$$p(y_i = +1) \ = \sigma(\alpha \sum_j \pi_{ij} \times y_j + \beta) \tag{9}$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is used to boost the probability of the rare tags. α and β stand for the two parameters to estimate the weights based on the distance for each word. For Eq. (9), the calculation is based on the probability of the presence of the label y in the image i. As can be observed, the labels y_{ca} and y_{cb} in one picture c, have the same transfer probabilities for the target image. The previous nearest-neighbor method such as [1] does not distinguish the importance of different labels in an image. This paper proposes a novel method with a new and more sophisticated model which distinguishes between the label weights while taking into account the image segmentation in addition to distance-based weights. The prediction of this method is defined as:

$$p(y_i = +1) \ = \sigma(\alpha \sum_j \pi_{ij} \times v(j, y_j) + \beta) \tag{10}$$

where $v(j, y_j)$ donates the weight of the label y in the image j based on the image segmentation. In order to extract local weights of labels, the “Frequency-tuned

Salient Region Detection (FSRD)" method [16], co-segmentation method [17] and Wordnet[2] are used to extract the area of each label. Algorithm 1 shows the whole algorithm for image annotation given by the LWTP method.

Algorithm 1. Image annotation

```
Input: Training set $T$ with semantic annotation set $V$ and unlabeled image $I$;
Output: Annotation results set $O = c1, c2, c3, c4, c5$
for the training set do
    use the metric learning to learn $\pi_{i,j}$ based on the distances
end for
for image $I$ do
    calculate the K nearest images
    for image $k$ from 1 to $K$ do
        for each word $y_j$ of image $k$ do
            calculate $v(k, y_j)$of the image K based on segmentation
        end for
    end for
    for each word $c$ do
        calculate $p(y_{ic} = +1)$
    end for
    obtain $c$ with 5 biggest $p(y_{ic} = +1)$
end for
```

4 Experiments and Results

4.1 Dataset and Evaluation Standards

In order to investigate the feasibility and the effectiveness of the proposed method, experiments are conducted on the Corel5K dataset. The Corel5K dataset contains 5,000 images from 50 Stock Photo CDs. Each CD includes 100 images and belongs to a particular theme. In the experiments, the visual feature set consists of global features Gist, RGB, LAB, HSV, local SIFT features, and Curvelet spectrum features. In order to evaluate the performance of the prediction annotation given by the LWTP method, each concept as a keyword is used to perform image retrieval operations on the dataset. The experiments take the average accuracy rate (Precision, P) and the average recall rate (Recall, R) over all concepts as evaluation metrics. At the same time, the number of annotation concepts with the positive recall value (N+) and the F1-measure (F1) are also considered in the experiments.

[2] http://wordnet.princeton.edu/.

4.2 Results and Analysis

Figure 2 shows the results of four examples. It includes four different types of pictures of the dataset where the left column shows the original labels of the image and right column shows the labels with probabilities given by the proposed method. It can be seen that the new results have more related tags. Table 1 shows the comparison results of the proposed LWTP method and other methods, including CRM, CMRM, etc. From Table 1, it can be observed that the LWTP method performs better than the traditional probabilistic methods such as CRM and CMRM in terms of precision, recall and F1-measure. The proposed method also performs better than the multi-class or multi-annotated image labeling algorithms such as SML. It also outperforms similar labeling methods, such as JEC, based on image feature extraction and label proximity propagation. Compared to the original TagProp method, the LWTP method obtains marked improvements of 2% for precision, 1% for recall, 2% for F1-measure and obtains 15 more words with positive recall.

Figures 3 and 4 show the retrieval P and the retrieval R for different values of nearest-neighbors (K). In these two figures, the blue curve represents the

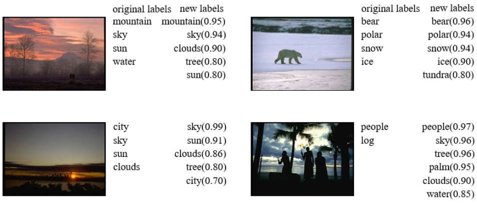

Fig. 2. Examples of the original and the new labels.

Table 1. Comparison of image auto-annotation effectiveness.

	P (%)	R (%)	F1-measure (%)	N+
CRM [6]	16	19	17.3	107
CMRM [5]	10	9	9.5	66
InfNet [18]	17	24	19.9	112
SML [8]	23	29	25.6	137
MBRM [7]	24	25	24.4	122
GS [19]	30	33	31.4	146
RF-optimize [20]	29	40	33.6	157
SVM-VT [9]	27	39	31.9	171
CNN-R [11]	32	41.3	37.2	166
JEC [10]	27	32	29.3	139
TagProp [1]	33	42	36.9	160
LWTP (our method)	**35**	**43**	**38.59**	**175**

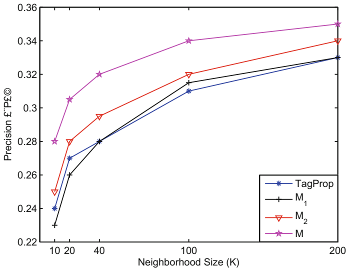

Fig. 3. Precision for different K.

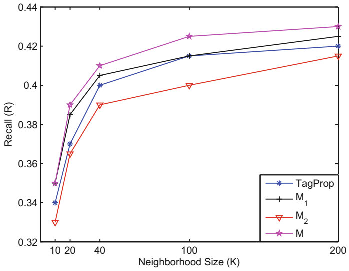

Fig. 4. Recall for different K.

results of the original TagProp method, and the pink curve marked with M represents the results of our method. This includes the dataset preprocessing and adjustment of the local tag weights based on image segmentation. The black curve represented by M1 (Method 1) is the experimental result of the original TagProp method with our label matrix preprocessing method. The red curve, indicated by M2 (Method 2) is an improved version of the TagProp method which does not carry out preprocessing but only adjusts the weights based on image segmentation. From the comparison between M1 and TagProp, it can observed that the tag library preprocessing step contribute more to R. Comparing the curves of M2 and the TagProp results, it can be seen that adjusting only the weights of the image tags contributes more to P. When M1 and M2 are combined, the effect of the combined method M is better than that of the TagProp method. In addition, if K is too small, the performance is relatively low and unstable. When the number of K reaches 200, the performance of the algorithm achieves the desired effect.

Figure 5 compares the Mean Average Precision (MAP) of the TagProp method and the proposed method. This result is also in line with the expected improvement. In order to compare the algorithm and supplement the experimental results, the co-segmentation method is also tried for image segmentation and object extraction. Figure 6 shows the P-R value for different SNs, where SN represents the number of similar images used for collaborative segmentation.

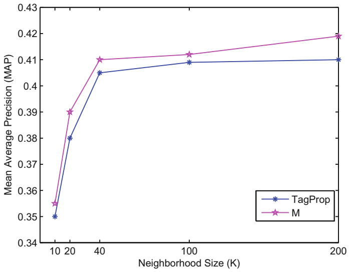

Fig. 5. Retrieval MAP for different K.

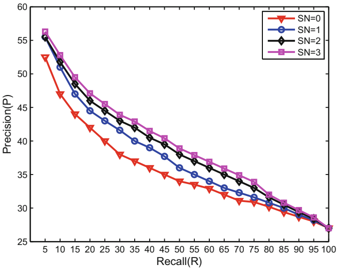

Fig. 6. Co-segmentation results.

As the value of SN increases, the P-R value also increases. However, the time complexity will also increase rapidly with increase in SN. Since there is no significant improvement in the P-R value for an increase of SN beyond 3, it is concluded that a value of SN less than 3 is preferable when using co-segmentation.

5 Conclusion and Future Work

In this paper, a novel method called LWTP was proposed for the automatic annotation of images. This method took into account both the content of images and the relations between an image and its neighboring images. By preprocessing the label matrix of the training set, this method balanced the differences in the distribution of the tags with low and high frequencies. Experimental results showed that the proposed method achieved an improved accuracy for image annotation, outperformed many baseline systems, and reached the state-of-art systems. In future works, we plan to adopt an optimization step to decrease the time complexity when preprocessing the label matrix. This study considers only the area when distinguishing the weights of labels based on image segmentation, and more factors will be investigated in future works.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (61572060, 61772060) and CERNET Innovation Project (NGII20151004, NGII20160316).

References

1. Mensink, T., Verbeek, J., Schmid, C., Guillaumin, M.: TagProp: discriminative metric learning in nearest neighbor models for image auto-annotation. In: IEEE International Conference on Computer Vision, pp. 309–316 (2010)
2. Yang, J.Y., Liu, G.H.: Content-based image retrieval using color difference histogram. Pattern Recogn. **46**(1), 188–198 (2013)
3. Chang, S.F., Kozintsev, I.V., Kennedy, L.S.: To search or to label?: predicting the performance of search-based automatic image classifiers. In: ACM International Workshop on Multimedia, Information Retrieval, pp. 249–258 (2006)
4. Shimada, A., Nagahara, H., Taniguchi, R.I., Xu, X.: Learning multi-task local metrics for image annotation. Multimed. Tools Appl. **75**(4), 1–29 (2014)
5. Lavrenko, V., Manmatha, R., Jeon, J.: Automatic image annotation and retrieval using cross-media relevance models. In: International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 119–126 (2003)
6. Manmatha, R., Jeon, J., Lavrenko, V.: A model for learning the semantics of pictures. In: NIPS, pp. 553–560 (2003)
7. Manmatha, R., Lavrenko, V., Feng, S.L.: Multiple Bernoulli relevance models for image and video annotation. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, vol. 2, pp. II-1002–II-1009 (2004)
8. Chan, A.B., Moreno, P.J., Vasconcelos, N., Carneiro, G.: Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **29**(3), 394–410 (2007)

9. Jawahar, C.V., Verma, Y.: Exploring SVM for image annotation in presence of confusing labels. In: British Machine Vision Conference, pp. 25.1–25.11 (2013)
10. Pavlovic, V., Kumar, S., Makadia, A.: Baselines for image annotation. Int. J. Comput. Vis. **90**(1), 88–105 (2010)
11. Maji, S., Manmatha, R., Murthy, V.N.: Automatic image annotation using deep learning representations. In: ACM on International Conference on Multimedia Retrieval, pp. 603–606 (2015)
12. Wang, Z., et al.: Weakly semi-supervised deep learning for multi-label image annotation. IEEE Trans. Big Data **1**(3), 109–122 (2015)
13. Chandra Sekhar, C., Sarangi, N.: Automatic image annotation using convex deep learning models. In: International Conference on Pattern Recognition Applications and Methods, pp. 92–99 (2015)
14. Wang, X.J., Zhang, L., Jing, F., Ma, W.Y.: AnnoSearch: image auto-annotation by search. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1483–1490 (2006)
15. Zhang, L., Wang, X.-J., Ma, W.-Y., Li, X.: Annotating images by mining image search results. IEEE Trans. Pattern Anal. Mach. Intell. **30**(11), 1919 (2008)
16. Hemami, S., Estrada, F., Susstrunk, S., Achanta, R.: Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1597–1604 (2009)
17. Irani, M., Faktor, A.: Co-segmentation by composition. In: IEEE International Conference on Computer Vision, pp. 1297–1304 (2014)
18. Metzler, D., Manmatha, R.: An inference network approach to image retrieval. In: Enser, P., Kompatsiaris, Y., O'Connor, N.E., Smeaton, A.F., Smeulders, A.W.M. (eds.) CIVR 2004. LNCS, vol. 3115, pp. 42–50. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-27814-6_9
19. Huang, J., Li, H., Metaxas, D.N., Zhang, S.: Automatic image annotation and retrieval using group sparsity. IEEE Trans. Syst. Man Cybern. Part B Cybern. Publ. IEEE Syst. Man Cybern. Soc. **42**(3), 838 (2012)
20. Fu, H., Zhang, Q., Qiu, G.: Random forest for image annotation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 86–99. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33783-3_7