Automatic Image Annotation Using Modified Keywords Transfer Mechanism Base on Image-Keyword Graph

Guo-Qing Xu[1], Zhi-Chun Mu[2]

[1] School of Automation and Electrical Engineering, University of Science and Technology Beijing
Beijing, 100083, China

[2] School of Automation and Electrical Engineering, University of Science and Technology Beijing
Beijing, 100083, China

Abstract

Automatic image annotation is widely considered to be an important yet open problem due to the well-known semantic gap. Recent works show that nearest-neighbor-based annotation approaches are simple and effective. In this paper we use a modified keywords transfer mechanism base on image-keyword unidirectional graph to derive a great annotator. The unidirectional graph describes the relationships between images and keywords, and can be derived from images with human annotations. On that basis, a modified keywords transfer mechanism base on visual neighbors is used to annotate new images. Our method achieves better annotation performance than two of the most advanced annotation methods in terms of precision, recall and F1 metrics on the open benchmark database.

Keywords: *Automatic Image Annotation, Semantic Gap, Keywords Transfer, Graph, Visual Neighbors.*

1. Introduction

Automatic image annotation is an important problem in many areas, such as image understanding and content-based image retrieval (CBIR) [1, 2]. The aim of image annotation techniques is to automatically assign a few relevant text keywords to test images, which reflect their visual contents [3, 4]. Automatic image annotation is a challenging task mainly for the reason that low-level features of images can hardly describe their high-level semantic concepts [5-7]. To solve this problem, considerable efforts have been made to improve the performance of image annotation in the past decade, and a number of image annotation algorithms have been proposed to explore semantic concept of images [1, 3, 8-12]. And image annotation has become a hot topic.

Generally, frameworks for automatic image annotation begin with representing images with low level features, and in the next train models for the semantic concepts from a number of semantically labeled image samples. And then the learned models are used to annotate new images. Many works focus on two main aspects of automatic image annotation, i.e. low level feature extraction and semantic model training, to bridge the semantic gap existing between low level features and high level semantics. From the view of semantic model, these methods can be broadly classified into three categories [3, 5]: generative models, discriminative models, and nearest-neighbor-based methods.

The generative models contain topic models and mixture models, and can be constructed by estimating the probability distribution over image features and high-level semantic concepts. Continuous Relevance Model[13], probabilistic latent semantic analysis[14], hierarchical Dirichlet process model [15] are typical generative models. Discriminative models treat each annotated keyword as an independent class, and for every keyword a separate classifier, for example the support vector machine classifier[16] or Gaussian Mixture Model[17], is learned. These classifiers are then used to predict which class the new image belongs to. The third type is nearest-neighbor-based methods. These approaches assume that semantic-relevant images share similar visual features and treat annotation as a retrieval problem.

Recently, the nearest-neighbor-based methods have captured more attentions from researchers due to its good annotation performance and simplicity in principle. Ma et al. [12] build a unified graph containing images and tags using the nearest neighbor searching method to bridge the semantic gap between image contents and tags. For one new image, top k nearest images was linked to it, and then a random walk model was carried out on the new graph to conduct image annotation. However, only global features were used in their method. Makadia et al. [8] proposed a simple nearest-neighbor tag transfer mechanism, and showed the state-of-the-art annotation performance. In their method, nearest neighbors are determined by averaging several normalized distances derived from different low level features. However, in most cases, the annotations of the new image are determined by the first nearest neighbor, and additional neighbors have no effect on the annotation results. Li et al. [18] propose a neighbor voting algorithm which learns tag relevance by accumulating votes from visual neighbors. Both the work in [8] and [18] can be regarded as weighted nearest neighbor voting model [7]. In this model, for each keyword, a seed image will receive relevance votes from its visual neighbors which are labeled with this keyword by users and the votes can be weighted according to their visual similarities. And in this situation, it is extremely important to choose a criteria to define which images are neighbors. In [8], Makadia used linear combinations of basic distances to yield the composite distance measure, namely Joint Equal Contribution (JEC) method. However, it is difficult to determine which distance measure is suitable for a certain feature because on various occasions there are several features and distance measures. Furthermore, usually more than one keyword is assigned to one image which is used as the ground truth, and the importance of these keywords to the image is different. As shown in Fig. 1, the keyword 'chair' is assigned five times and 'person' is assigned three times and it is reasonable to believe that 'chair' is more relevant to this image.

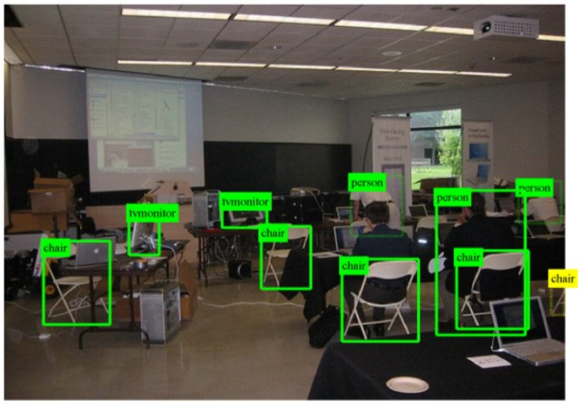

Fig. 1 example image with its annotations.

In this paper we present an improved nearest-neighbor-based method, which was inspired by the previous study [8, 12]. The proposed approach fuses a simple image-keyword unidirectional graph and a modified keywords transfer mechanism, and achieves better annotation performance than two of existing well-known methods in terms of precision, recall and F1 on the open benchmark dataset.

2. Proposed Approach

In the proposed approach, we will assign a few given keywords to test image if most of its neighbor images share the given keywords. To this end, we first build an image-keyword unidirectional graph to capture the relationships between images and keywords. And then we represent images with six low level features, which are used to find the nearest-neighbors of test image. In the next, we present a modified keywords transfer mechanism that employs visual similarity and image-keyword unidirectional graph to annotate new image. We detail our approach in the following.

2.1 Image-Keyword Unidirectional Graph

We build the image-keyword unidirectional graph to show the relationships between images and keywords. As indicated in the introduction, the correlations between an image and different keywords are not necessarily the same. Let D denote the annotated image set, and D={Im i |i=1…p}, where p is the total number of annotated images. The keyword set is denoted by W, and W={Keyword j| j=1…q}. Fig.2 indicates the image-keyword unidirectional graph.

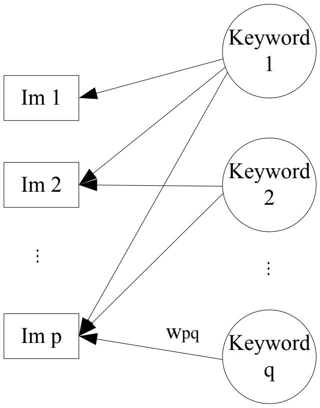

Fig. 2 image-keyword unidirectional graph

As shown in Fig.2, the left nodes of the unidirectional graph denote the images, and the right nodes represent the keywords. We use y_{ij} to encode the image annotations, and its value is determined by the times of the keyword j assigned for image i. For a given keyword q and an image p, if the corresponding annotation $y_{pq}>0$, then there is a directed image-keyword edge between Im p and keyword q. The weight of this directed edge is computed by:

$$w_{pq} = y_{pq} / \sum_{i=1}^{q} y_{pi} \qquad (1)$$

2.2 Visual Feature Extracting

Extracting effective visual features from image pixels is an important part in image understanding and content-based image retrieval. Many feature extraction techniques have been proposed in the literature. Commonly used visual features in automatic image annotation include color, texture and shape. To comprehensively describe the images and maximize the amount of information extracted, six low level features with discriminating power are extracted in total, four of which are MPEG-7 visual descriptors. They are Scalable Color descriptor, Color Layout descriptor, Grid Color Moment, Homogeneous Texture descriptor, Edge Histogram descriptor, and Wavelet Moment.

Scalable Color descriptor is derived by applying Haar transform on a color histogram in the HSV color space. The dimensionality of the Descriptor used here is 256. Color Layout descriptor captures the spatial distribution of color and is based on coefficients of the Discrete Cosine Transform. This descriptor is very compact. Grid Color Moment is an 81-dimensional vector and derived on nine equal-sized grids for each image. For each grid, three kinds of color moments in each of three color channels are extracted. Homogeneous Texture descriptor describes the texture information using the mean and standard deviation in the frequency domain of an image filtered by Gabor functions. Additionally, the mean and standard deviation of the original image are also contained in the descriptor. Edge Histogram descriptor is represented by a histogram with eighty bins and describes five types of edges in an image, including vertical, horizontal, 45-degree diagonal, 135-degree diagonal and non-directional edges. Wavelet Moment simultaneously captures the spatially and local features, which is constructed using Cubic B-spline wavelets as the basis functions in this paper.

2.3 Modified Keywords Transfer Mechanism Base on the Image- Keyword Unidirectional Graph

For a new image I_m to be annotated, first we compute visual similarities between I_m and the annotated images to find its nearest neighbors. Let Ii indicate the i-th image in the annotated image set D, and its six feature vectors are represented by Fi = [f_{1i}, f_{2i} . . . , f_{6i}]. We employ the city block function to measure the dissimilarity between two images I_i and I_m. Similar to the JEC method, we compute distance between corresponding features f_{ki} and f_{kj} in two images I_i and I_m.

$$dis_k(I_i, I_m) = \sum(| f_{ki} - f_{km} |) \tag{2}$$

Where dis_k means the distance between the k-th feature vectors of the two images, k =1, 2… 6. Then for each feature, the distance is normalized using the maximum. And the total distance between I_i and I_m is the sum of the scaled distance between corresponding features. The smaller the distance is, the greater the similarity between the two images is.

Based on the calculated image similarities, the nearest neighbors of image I_m can be determined by ranking the similarities in descending order. The top N most similar images are used to propagate the annotation keywords to the new image. Meanwhile, the similarities between I_m and these images are also utilized to weight the measure the degree of relationship between them. Suppose that I_{m1}, I_{m2}, …,I_{mN} are the top N most similar images, and the similarities between them and I_m are sim_{m1}, sim_{m2},…sim_{mN}, respectively. For a given keyword q in W, the probability that keyword q will be assigned to I_m can be calculated by:

$$prob(q / \mathrm{I_m}) = \sum_{i=1}^{N} sim_{mi} \times w_{iq} \tag{3}$$

For image I_m, the probabilities of each keyword in W are sorted in descending order and the top ranked keywords are assigned to I_m as the annotations.

3. Experiments and Discussion

3.1 Experimental Data and Performance Metrics

Our experiments test the annotation performance of the proposed method for image annotation on the PascalVOC07 (Pascal Visual Object Classes Challenge 2007) database [19]. The PascalVOC07 database is a standard dataset of images and annotation, and contains 9963 annotated images in total. 8967 images are adopted as the ground truth and 996 images are used as the testing set, with the test-train ratio similar to that in [8]. There are 20 unique concepts in this database such as bicycle and bird, and on average an image contains 2.47 keywords.

To measure the effectiveness of our proposed method, three measures, namely precision, recall and F1, are used in the experiments following previous studies [8, 20]. For a given keyword j in keyword set W, in the testing set suppose the number of human annotated images with this keyword is N_{Gj}, the number of images annotated with this keyword by our method is N_{Mj}, where the number of correct annotations is N_{Cj}. The recall, precision and F1 of this keyword are calculated as

$$Precision_j = N_{Cj} / N_{Mj} \tag{4}$$

$$Recall_j = N_{Cj} / N_{Gj} \tag{5}$$

$$F1_j = \frac{2 \times Precision_j \times Recall_j}{Precision_j + Recall_j} \tag{6}$$

The recall, precision and F1 values shown in the following are obtained by averaging the results over all the keywords. One thing to note here is that the definitions of precision and recall [8, 17, 20] here are different from that in [5].

3.2 Results and Comparisons

When testing the annotation performance of our method, dissimilar to other approaches, we vary the number of keywords assigned to each image from 2 to 6 for this experiment. And the proposed method was compared with two other notable annotation algorithms, including AICDM (Annotation by Image-Concept Distribution Model) [9], LT (Label Transfer mechanism) [8] in terms of precision, recall and F1 metrics. AICDM was proposed to facilitate image annotation by discovering the associations between visual features and human concepts. Their empirical results showed that this method outperformed the Continuous Relevance Model and SVM method. Label Transfer mechanism is a simple yet powerful annotation method, which transfer keywords to a new image from its nearest neighbors with ground truth annotations. We made our best effort to implement those algorithms based on their papers and compare their algorithm with ours.

The number of nearest neighbors influences probability that a keyword will be assigned to test image, hence it is an important parameter in the proposed method. Firstly we test our method with different nearest neighbors for image annotation. Table 1, 2and 3 show the effect of varying the number of nearest neighbors between 5 and 30 with an interval of five. From table 1, we can see that the recall increases with the increasing of numbers of nearest neighbors when the number of assigned keywords is relatively large, since there are more correct annotations in the candidate set as more neighbor images are included.

Table 1: recall values under different nearest neighbors (%)

No. of nearest neighbors	2 words	3 words	4 words	5 words	6 words
5	35.23	42.17	49.16	54.37	56.24
10	33.34	43.52	50.23	55.79	61.24
15	32.22	42.50	51.35	56.77	61.89
20	30.40	41.51	51.86	59.39	65.25
25	31.29	41.73	50.91	59.46	65.80
30	31.23	41.75	52.12	59.70	65.94

Table 2: precision values under different nearest neighbors (%)

No. of nearest neighbors	2 words	3 words	4 words	5 words	6 words
5	24.54	19.30	17.26	16.33	15.69
10	25.08	21.07	17.53	15.48	14.12
15	25.45	21.05	18.22	15.65	14.01
20	24.78	21.27	18.64	16.35	14.78
25	26.98	21.44	18.53	16.50	14.92
30	27.81	21.45	19.20	16.83	15.04

As shown in table 2, for a certain number of keywords assigned to one image, the precision values maintain relative stability when the neighbor size is changed, which indicates the robustness of the proposed method. According to table 3, we observe that it is appropriate in the proposed method to set the number of nearest neighbors to 30 for image annotation.

Table 3: F1 values under different nearest neighbors (%)

No. of nearest neighbors	2 words	3 words	4 words	5 words	6 words
5	28.93	26.48	25.55	25.12	24.53
10	28.63	28.39	25.99	24.23	22.95
15	28.44	28.16	26.90	24.53	22.85
20	27.30	28.13	27.42	25.64	24.10
25	28.97	28.33	27.18	25.83	24.33
30	29.42	28.34	28.06	26.26	24.49

We also find that as the number of returned keywords grows from 2 to 6, the recall increases while the precision decreases. And similar phenomenon occurs frequently in image annotation and retrieval [11].

Then we compare the annotation performance of the proposed method with that of AICDM and LT. For AICDM, the employed visual features are Scalable Color descriptor and Homogeneous Texture descriptor, following [9]. And for LT, all six visual features are used. Fig 3 and 4 show how the recall, precision and F1 values of the three methods change over the number of assigned keywords, where our method is denoted by MKT. We can obtain some important observations. First, both LT and MKT methods achieve much better results than AICDM in terms of recall, precision and F1, mainly because the parameter settings of AICDM is not optimal and determined manually due to lack of theory or experiment instructions in their paper[9]. Second, MKT obtains higher recall and precision values that LT method in most cases, which reveals that our method is more effective.

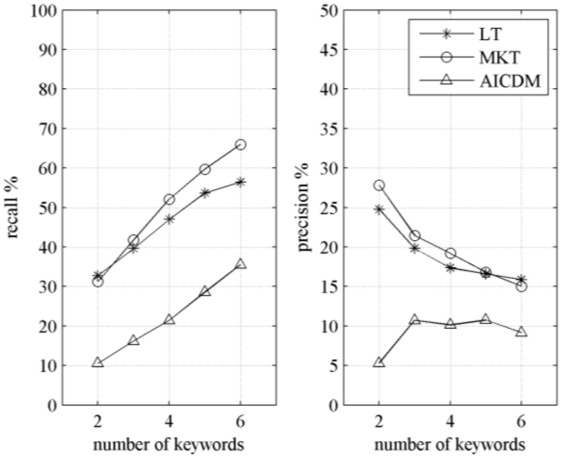

Fig. 3 recall and precision comparisons in image annotation

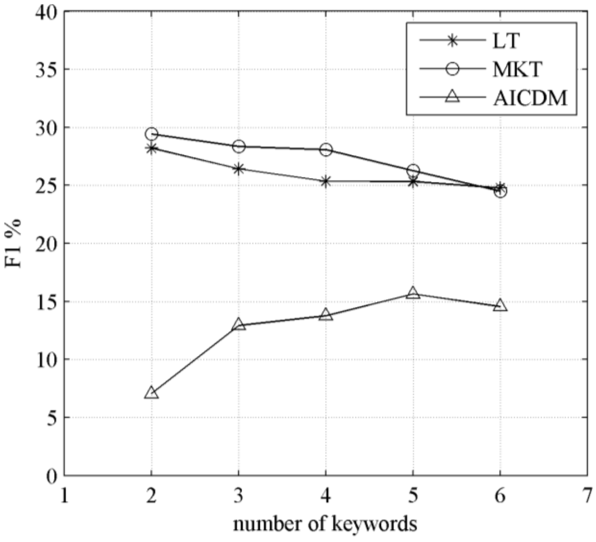

Fig. 4 F1 comparisons in image annotation

4. Conclusions

Automatic image annotation is widely considered to be an open issue which is hard to resolve due to the so-called semantic gap. Inspired by previous works, we fuse a modified keywords transfer mechanism and an image-keyword unidirectional graph to annotate new images. The unidirectional graph describes the relationships between images and keywords. And a modified keywords transfer mechanism base on the unidirectional graph is derived. Our method achieves better annotation performance than two of the most well-known methods in terms of precision, recall and F1 on the PascalVOC07 database.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.61170116).

References

[1] D. Zhang, M.M. Islam, G. Lu, A review on automatic image annotation techniques, Pattern Recognition, Vol.45,No.1, 2012,pp. 346-362.

[2] V. Khanaa, M. Rajani, K.A.A. Raj, Efficient Use of Semantic Annotation in Content Based Image Retrieval (CBIR), International Journal of Computer Science Issues, Vol.9,No.2 2-2, 2012,pp. 273-279.

[3] Z. Shaoting, H. Junzhou, L. Hongsheng, D.N. Metaxas, Automatic Image Annotation and Retrieval Using Group Sparsity, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, Vol.42,No.3, 2012,pp. 838-849.

[4] M.B.M. Abdelaziz, Z. Lynda, Improving automatic image annotation: Approach by bag-of-key point, International Journal of Computer Science Issues, Vol.9,No.5 5-1, 2012,pp. 323-328.

[5] S. Ja-Hwung, C. Chien-Li, L. Ching-Yung, V.S. Tseng, Effective Semantic Annotation by Image-to-Concept Distribution Model, Multimedia, IEEE Transactions on, Vol.13,No.3, 2011,pp. 530-538.

[6] R. Datta, D. Joshi, J. Li, J.Z. Wang, Image retrieval: Ideas, influences, and trends of the new age, ACM Computing Surveys, Vol.40,No.2, 2008,pp. 1-60.

[7] Z. Li, M. Lew, Cost-sensitive learning in social image tagging: review, new ideas and evaluation, International Journal of Multimedia Information Retrieval, Vol.1,No.4, 2012,pp. 205-222.

[8] A. Makadia, V. Pavlovic, S. Kumar, Baselines for image annotation, International Journal of Computer Vision, Vol.90,No.1, 2010,pp. 88-105.

[9] S. Ja-Hwung, C. Chien-Li, L. Ching-Yung, V.S. Tseng, Effective image semantic annotation by discovering visual-concept associations from image-concept distribution model, in: Multimedia and Expo (ICME), 2010 IEEE International Conference on, 2010, pp. 42-47.

[10] M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid, TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation, in: Computer Vision, 2009 IEEE 12th International Conference on, 2009, pp. 309-316.

[11] L. Wu, R. Jin, A. Jain, Tag Completion for Image Retrieval, Pattern Analysis and Machine Intelligence, IEEE Transactions on, Vol.pp,No.99, 2012,pp. 1-1.

[12] M. Hao, Z. Jianke, M.R.T. Lyu, I. King, Bridging the Semantic Gap Between Image Contents and Tags, Multimedia, IEEE Transactions on, Vol.12,No.5, 2010,pp. 462-473.

[13] R. Jin, J.Y. Chai, L. Si, Effective automatic image annotation via a coherent language model and active learning, in: Proceedings of the 12th annual ACM international conference on Multimedia, ACM, New York, NY, USA, 2004, pp. 892-899.

[14] F. Monay, D. Gatica-Perez, PLSA-based image auto-annotation: constraining the latent space, in: Proceedings of the 12th annual ACM international conference on Multimedia, ACM, New York, NY, USA, 2004, pp. 348-351.

[15] O. Yakhnenko, V. Honavar, Annotating images and image objects using a hierarchical dirichlet process model, in: Proceedings of the 9th International Workshop on Multimedia Data Mining: held in conjunction with the ACM SIGKDD 2008, ACM, Las Vegas, Nevada, 2008, pp. 1-7.

[16] X. Qi, Y. Han, Incorporating multiple SVMs for automatic image annotation, Pattern Recognition, Vol.40,No.2, 2007,pp. 728-741.

[17] G. Carneiro, A.B. Chan, P.J. Moreno, N. Vasconcelos, Supervised Learning of Semantic Classes for Image Annotation and Retrieval, Pattern Analysis and Machine Intelligence, IEEE Transactions on, Vol.29,No.3, 2007,pp. 394-410.

[18] L. Xirong, C.G.M. Snoek, M. Worring, Learning Social Tag Relevance by Neighbor Voting, Multimedia, IEEE Transactions on, Vol.11,No.7, 2009,pp. 1310-1322.

[19] M. Everingham, L. Gool, C.K. Williams, J. Winn, A. Zisserman, The Pascal Visual Object Classes (VOC) Challenge, International Journal of Computer Vision, Vol.88,No.2, 2010,pp. 303-338.

[20] M. Wang, F. Li, M. Wang, Collaborative visual modeling for automatic image annotation via sparse model coding, Neurocomputing, Vol.95,No.2012,pp. 22-28.

Guo-Qing Xu received the B.E. degree from the School of Electrical Engineering at Zhengzhou University in 2008. He is currently pursuing Ph.D. in Control Science and Control Engineering from the School of Automation and Electrical Engineering, University of Science and Technology Beijing. His research interests include content-based image retrieval, automatic image annotation, machine learning, and pattern recognition.

Zhi-Chun Mu received the B.E. and M.E. degrees from the Department of Automation at University of Science and Technology Beijing in 1978 and 1983, respectively. He is currently a professor in the School of Automation and Electrical Engineering, University of Science and Technology Beijing. He was a visiting scholar at Davy International UK and Sheffield Polytechnic, England, from 1989 to 1991, and a visiting Professor at University of Brighton, England from 1996 to 1999. As a Guest Professor, he visited Laboratoire d'Electronique, d'Informatique et d'Image, CNRS University of Burgundy France in 2007. He was the Chair of Organizing Committee of IEEE International Conference on Wavelet Analysis and Pattern Recognition 2007. He has served as a reviewer and member of Evaluation and Assessment Commission, Division of Information Science, National Natural Science Foundation of China since 2002.He is now Chair of IEEE SMC Beijing (capital region) Chapter. His main research interests include Pattern Recognition and Biometrics, Artificial Intelligence and its applications, Data Mining as well as Process Control and Modeling. He has published more than 180 refereed journal and conference papers in these areas. He is the corresponding author of this paper.