Proceedings of the 2012 International Conference on Machine Learning and Cybernetics, Xian, 15-17 July, 2012

SEMI-AUTOMATIC IMAGE ANNOTATION USING SPARSE CODING

WEIFENG ZHANG[1], ZENGCHANG QIN[1], TAO WAN[2]

[1] Intelligent Computing and Machine Learning Lab
School of ASEE, Beihang University, Beijing, 100191, China
[2] School of Medicine, Boston University, Boston, MA 02215, USA
E-MAIL: zcqin@buaa.edu.cn

Abstract:

Automatically assigning keywords to images is of great interest as it allows one to index, retrieve, and understand large collections of image data. It has become a new research focus and many techniques have been proposed to solve this problem. In this paper, a novel semi-auto image annotation technique is proposed. The new developed method uses a label transfer mechanism to automatically recommend promising tags to each image by assigning each image a category label first. Since image representation is one of the key problems in image annotation, we utilize a sparse coding based spatial pyramid matching as an effective way to model and interpret image features. Experimental results demonstrate that the proposed method outperforms the current state-of-the-art methods on two benchmark image datasets.

Keywords:

Image annotation, Sparse coding, Spatial pyramid matching, Bag-of-features.

1. Introduction

Users can have a better understanding of the semantic content of multimedia data aided by annotation. However, manually annotating images requires time and effort, and it is difficult for users to provide all relevant tags for each image. Thus automatic image annotation emerged and has recently attracted lots of attention. The goal of automatically annotating a image is to assign a few relevant text words to the image to reflect its semantic content. It allows one to improve the quality of image search by utilizing image content to a set of fast indexed and searchable keywords. In recent years, there have been a number of techniques developed, such as [1, 2, 3, 6, 11]. Most of these methods define either parametric or non-parametric machine learning models to capture the latent relationship between the image features and keywords. For example, cross media relevance model (CMRM) [3], continuous relevance model (CRM) [6], and multiple Bernoulli relevance model (MBRM) [2] assume different, non-parametric density representations of the joint word-image space.

Even though some of these classical techniques have shown impressive results, they employ a global learning scheme by considering overall images in the training set. Therefore, to get appropriate semantic tags for a given test image, one need to compare this image with every image in the training set and then propagate the tagging information of the training images to this image. Obviously, this type of global learning approach is time-consuming because all the training images will be processed during the annotation procedure. In this paper, we propose a novel local learning method based on image category tags to perform semi-automatic image annotation. The proposed method requires users to assign each image with one category label first in order to remove those semantically irrelevant images and the remaining relevant images are used for tag propagation. The appropriate use of human prior knowledge will significantly improve the tagging performance and build a reliable annotation system. Image representation is a critical issue and we solve this problem by using sparse coding based spatial pyramid matching to model scale-invariant feature transform (SIFT) image features. Experimental results show that the recommended keywords can effectively reflect the image content.

2. Image Representation

Image representation is a key issue in image annotation. Bag-of-features (BoF) and spatial pyramid matching (SPM) [7] are two popular state-of-the-art image representations. Although BoF and SPM work well for image classification, we use sparse coding based spatial pyramid matching (ScSPM) method [12] to model image features. A typical ScSPM scheme mainly consists of four steps: 1) feature points are detected within the input image, and descriptors such as "SIFT" or

978-1-4673-1487-9/12/$31.00 ©2012 IEEE

"color moment" are extracted from each feature point; 2) a codebook is generated by using sparse coding; 3) multiple codes for each sub-region are integrated by averaging and normalizing into a histogram; 4) the histograms from all sub-regions are concatenated together to form a final representation of the image.

Generally, given $X = [x_1, x_2, ..., x_N]$, $x_i \in \mathbb{R}^{d\times 1}$ as the data matrix of local feature space. Spare coding algorithm is exploited to learn the codebook as follows:

$$\min_{U,V} \sum_{i=1}^{N} ||x_i - Uv_i||^2 + \lambda||v_i||_1 \quad s.t. : |u_j| \leq 1, \forall j = 1, ..., k. \tag{1}$$

where v_i is the sparse codes for the local feature x_i, the matrix U is the codebook, and the parameter λ is used to control the sparsity. The conventional way for solving the above optimization problem is to alternatively optimize on V or U while fixing the other over numerous iterations. Lee *et al.* [8] designed an efficient sparse coding algorithm [1] to handle this problem in which the optimization is convex in V (with U fixed) and then convex in U (with V fixed), but not simultaneously. By fixing U, Eq.1 can be solved by optimizing over each coefficient v_m individually:

$$\min_{v_m} ||x_m - Uv_m|| + \lambda|v_m| \tag{2}$$

This is essentially a linear regression problem with L_1 norm regularization on the coefficients that is well known as Lasso in statistical literatures. Then fixing V, it reduces to a least square problem with quadratic constraints:

$$\min_{U} ||X - UV||_F^2 \quad s.t. : ||v_k|| \leq 1, \forall k = 1, 2, ..., K. \tag{3}$$

The optimization can be achieved efficiently by the Lagrange dual as described in [8].

We are able to learn the codebook from the above sparse coding approach, in which each basis vector represents one basic local patch pattern, and hundreds of thousands of sparse codes. Each entry of certain sparse code represents the response of the patch to the corresponding basic pattern in the codebook. Following the work of Yang *et al.* [12], a maximum pooling based image representation is used. Suppose one image region has n local features, and the codebook size is K. The sparse codes for these local features are $[v_1, v_2, ..., v_n]$. After the maximum pooling, each image will be represented by a K dimensional vector y, and the l^{th} entry is the largest response to the l^{th} basis vector of the all the sparse codes in the selected region: $y_l = \max\{|v_{1l}|, |v_{2l}|, ..., |v_{nl}|\}$, where v_{nl} is the l^{th} entry v_n.

[1] The source code of this algorithm is available from *http://www.eecs.umich.edu/~honglak/softwares/nips06-sparsecoding.htm*

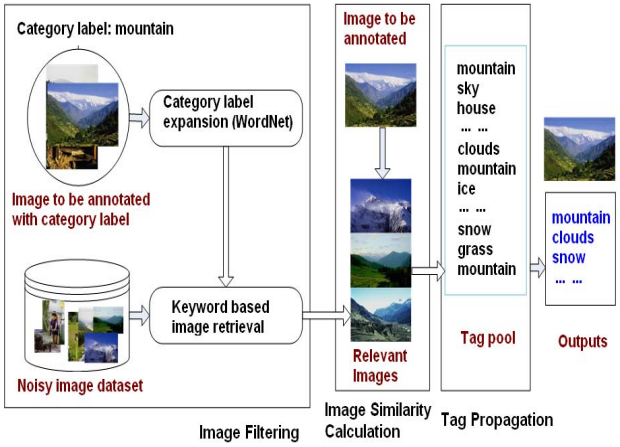

Figure 1. Flowchart of sparse coding based semi-automatic image annotation system given category tags.

3. Image Annotation

Our proposed method contains three main modules: (i) Image filtering: noisy images are filtered out by using the testing image's category label as the keyword. (ii) Calculating image similarity based on visual features: a sparse coding based spatial pyramid matching is used to model visual features. (iii) Tag propagation: tags corresponding to the relevant images are transferred to the testing image based on the statistical relationship.

3.1. Noisy Image Filtering

Images in a large training data set are relatively noisy, which means that lots of them are semantically irrelevant to the image to be annotated. Using these "noisy" images directly might deteriorate the annotation performance, so it is necessary to remove these irrelevant images. Different from classic automatic image annotation systems, our approach first assigns each image with a particular category label to be served as a keyword. Thus traditional image retrieval method based on keyword can be used to filter out the noisy images.

In this work we use WordNet to expand the given category label. WordNet is a lexical database for the English language.

It groups English words into sets of synonyms called synsets, which provides short, general definitions, and records the various semantic relations between these synonym sets. The reasons for using WordNet are in the following aspects: (i) The category labels are provided by users, so different users might give different labels to the same image. However, it is reasonable to assume that these labels are semantically similar. (ii) The given category labels may not appear in the training images particularly when the size of training image set is small. In this case, there will be no image returned through the keyword based image retrieval using this category label as a keyword. (iii) One category label might not be sufficient to define the content of testing image. A proper category label expansion could bring several more words to better describe the testing image. Thereby, WordNet is used here to expand one category label to a few keywords.

3.2. Tag Propagation

The tag propagation module intends to propagate the tags of relevant images to the test images based on the joint probability of observing an image with possible annotation keywords. Given a relevant image set $\mathcal{T}$, the joint probability of an image $J \in \mathcal{T}$ and a word is defined by:

$$P(J, w) = P(w|J)P(J) \tag{4}$$

where w is a annotation assigned to image J. The joint probability is computed by two distinct probability distributions: $P(w|J)$ is the conditional probability of generating w given J, and $P(J)$ is the probability of selecting image J in the underlying model. Normally, it is considered as a uniform distribution.

Given a test image I, we assume that I comes from the same process that each training image $J \in \mathcal{T}$ has been generated. However, without knowing the annotation of I, we cannot decide which process it is. Hence, we compute an expectation over all images $J \in \mathcal{T}$, and the joint probability of a set of keywords W and a test image I is given by:

$$\begin{aligned} P(I, w) &= \sum_{J \in \mathcal{T}} P(I, w|J)P(J) \\ &= \sum_{J \in \mathcal{T}} P(w|J)P(I|J)P(J) \\ &= \sum_{J \in \mathcal{T}} P(w|J)P(I, J) \end{aligned} \tag{5}$$

where the test image I is assumed to be independent of W given a training image J. Although any kind of distribution can be used to model annotation keywords (i.e. $P(w|J)$) in the above relevance model, here we utilize the Laverenko's method [6]:

$$P(w|J) = \frac{\mu p_w + N_{w,J}}{\mu + \sum_{w'} N_{w',J}} \tag{6}$$

where $N_{w',J}$ is the number of times the keyword w appearing in the annotations of training image J. P_w is the frequency of keyword w in the entire training set, and μ is a parameter which can be obtained via the experiment. μ determines the degree of interpolation with a larger value of μ giving more precedent to the background probability over the image annotation word frequencies.

Finally, we need to solve the conditional distribution $P(I|J)$ that actually measures the similarity between the test image I and the training image J. As we mentioned in Section 2 that an image can be represented by a d dimensional vector y by using maximum pooling, the conditional probability is calculated by:

$$P(I|J) = \frac{similarity(I, J)}{\sum_{j=1}^{N} similarity(I, J_j)} \tag{7}$$

where $similarity(I, J)$ is obtained by the histogram intersection distance (HID) between two images. When image I and image J are represented as $[y_1^I, y_2^I, ..., y_d^I]$ and $[y_1^J, y_2^J, ..., y_d^J]$, respectively, the HID between them can be calculated as follows:

$$similarity(I, J) = \frac{\sum_{i=1}^{d} min(y_i^I, y_i^J)}{\sum_{i=1}^{d} y_i^I} \tag{8}$$

4. Experimental Results

In the following experiments, the proposed image annotation method is tested on two widely used benchmark image datasets. $Corel5k$ contains $5,000$ images from 50 Stock Photo CDs. Each image has 1 to 5 keywords making 374 keywords in total. It has been used by various methods for image annotation [1, 2, 3, 6]. In our experiments, $4,500$ out of $5,000$ images are used for the training data set and the rest 500 images are for the testing. The *ICPR2005* image database[2] is a small image dataset with 1109 images selected from 20 natural scenes, and each image is assigned with 1 to 22 keywords. The dataset includes totally 429 keywords among which 266 will appear in the test set. We randomly choose 300 images with 15 images from each natural scene, for the testing and the remaining images are for the training.

The following widely used measures are adopted to evaluate the annotation performance. *Recall* is computed as the number of images correctly annotated with a given word divided

[2] *http://www.cs.washington.edu/research/imagedatabase/groundtruth/*

	$\#(R_W > 0)$	$\overline{R}_n$	$\overline{P}_n$
Co-occ[10]	19	0.02	0.03
MT[1]	49	0.04	0.06
CMRM[3]	66	0.09	0.10
CRM[6]	107	0.19	0.16
FastAN[4]	-	0.09	0.06
MCML[5]	-	0.12	0.07
MBRM[2]	122	0.25	0.24
SMK+GRM[9]	143	0.334	0.301
Our Method	**225**	**0.66**	**0.57**

Table 1. Comparison of results on *Corel5k*.

	$\overline{R}_n$	$\overline{P}_n$
Co-occ[10]	0.02	0.02
CMRM[3]	0.19	0.18
CRM[6]	0.22	0.20
FastAN[4]	0.21	0.17
MCML[5]	0.14	0.11
Our Method	**0.48**	**0.31**

Table 2. Comparison of results on *ICPR 2005*

by the number of images that have that word in the human annotation. The metric measures the completeness in annotating images with word w: $R_w = \frac{c_w}{e_w}$ where c_w is the number of correct images annotated with word w, and e_w is the number of images annotated with w in the ground truth. *Precision* is defined as the number of correctly annotated images divided by the total number of images annotated with that particular word (correctly or not). This metric is used to measure the accuracy in annotating images with word w: $P_w = \frac{c_w}{r_w}$ where r_w is the number of images has been annotated with word w by the system. *Number of words with recall greater than zero*, denoted by $\#(R_w > 0)$, is a metric to evaluate the capability of the system labeling images with rare keywords. Usually, these images are difficult to annotate due to the fact that a small number of positive instances are available in the training set. We calculate the average of recall and precision over all the n words, and they are given by: $\overline{R}_n = \frac{1}{n}\sum_{w=1}^{n} r_w$, and $\overline{P}_n = \frac{1}{n}\sum_{w=1}^{n} p_w$.

We compare our proposed method with 8 reference models using different image representations, including Co-occ [10], MT [1], CMRM [3], CRM [6], FastAN[4], MCML[5], MBRM [2], SMK+GRM [9]. In this experiment, each image is annotated with 5 tags and then these tags are compared with the ground truth tags. The experimental results using the dataset $Corel5k$ are summarized in Table 1. Our method gains 0.66 and 0.57 for the average per-word recall and average per-word precision reach, respectively, while the best results for the reference methods are 0.334 and 0.301. Moreover, the proposed method yields 225 words with recall over zero which has 82 more words than the second best result of SMK+GRM model. The results demonstrate that our method achieves a superior performance in comparison with other state-of-the-art techniques. Our developed model has a number of advantages. In our method, each image is tagged with a category label given by user which can be considered as a strong prior knowledge. Although the test models used for comparison do not have such strong priors available, we still offer a feasible and robust approach to annotate images. Such examples include Microsoft's Lazy Snapping that uses strong human expert knowledge for image segmentation and object detection. In addition, The sparse coding based spatial pyramid matching is used to describe image features which leads to a better representation of the semantics of images.

Though the Corel set has served as a common evaluation platform for image annotation, it is often criticized for its bias due to appearance variance and contrived annotations. Therefore, we use another image set *ICPR2005* to provide a complementary view. The experimental results are listed in Table 2. The number of words with recall greater than zero is not displayed in Table 2. The reason is that no such records in the cited publications. Again we notice that our model significantly outperforms the current advanced models. Our average per-word recall reaches 0.48 which is 0.26 higher than CRM. Also the average per-word precision is 0.31 with 0.11 higher than CRM.

5. Conclusions

In this paper, we present a semi-automatic image annotation model based on a sparse coding representation of the images. To accurately and effectively represent the content of image, sparse coding based spatial pyramid matching technique is employed. In our new model, the image is given the category label by users which is different from traditional automatic image annotation approaches. Apart from manually assigning the category label, the proposed method performs fully automatically to annotate images. The human prior knowledge is well used to design a feasible and effective annotation system. The method have been evaluated using two well known benchmark data sets. The experiments show that the new proposed method greatly improves the annotation performance in different evaluation metrics.

Acknowledgements

This work is partially funded by the NCET Program of MOE, China and the SRF for ROCS.

References

[1] P. Duygulu, K. Barnard, J.F.G. de Freitas, and D.A. Forsyth. "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary". *ECCV*, pp. 97-112, 2002.

[2] S.L. Feng, R. Manmatha, and V. Lavrenko. "Multiple bernoulli relevance models for image and video annotation". *CVPR*, Vol. 2, pp. 1002-1009, 2004.

[3] J. Jeon, V. Lavreko, and R. Manmatha. "Automatic image annotation and retrieval using cross-media relevance models". In *ACM SIGIR Conf. Research and Development in Information Retrieval*, pp. 119-126, 2003.

[4] H. Kwasnicka, M. Paradowski. "Fast image auto-annotation with discretized feature distance measures". *Machine Graphics and Vision*, Vol. 15(2), pp. 123-140, 2006.

[5] H. Kwasnicka, M. Paradowski. "Multiple class machine learning approach for image auto-annotation problem". *ISDA*, pp. 347-352, 2006.

[6] V. Lavrenko, R. Manmatha, and J. Jeon. "A model for learning the semantics of pictures". *NIPS*, 2004.

[7] S. Lazebnik, C. Schmid, and J. Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories". *CVPR*, pp. 2169-2178, 2006.

[8] H. Lee, A. Battle, R. Raina, and A. Y. Ng. "Efficient sparse coding algorithms". *NIPS*, 2006.

[9] Z. Lu and H. H. Ip. "Generalized Relevance Models for automatic image annotation". *PCM*, pp. 245-255, 2009.

[10] Y. Mori, H.Takahashi, and R. Oka. "Image-to-word transformation based on dividing and vector quantizing images with words". *Intl. Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[11] L. Wang, L. Liu, and L. Khan. "Automatic image annotation and retrieval using subspace clustering algorithm". In *ACM MMDB*, pp. 100-108, 2004.

[12] J. Yang, K. Yu, Y. Gong, and T. Huang. "Linear spatial pyramid matching using sparse coding for image classification". *CVPR*, pp. 1794-1801, 2009.

[13] X. Yuan, J. Yu, Z. Qin and T. Wan. "A bag-of-features model with integrated SIFT-LBP features for content-based image retrieval", *ICIP*, pp. 1061-1064, 2011.