Collaborative and Content-based Image Labeling

Ning Zhou[1] William K. Cheung[2] Xiangyang Xue[1] Guoping Qiu[3]
[1]*School of Computer Science, Fudan University, Shanghai, China*
[2]*Dept. of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong*
[3]*School of Computer Science, The University of Nottingham, UK*

Abstract

Many on-line photo sharing systems allow users to tag their images so as to support semantic image search. In this paper, we study how one can take advantages of the already-tagged images to (semi-)automate the labeling of newly uploaded ones. In particular, we propose a hybrid approach for the prediction where user-provided tags and image visual contents are fused under a unified probabilistic framework. Kernel smoothing and collaborative filtering techniques are explored for improving the accuracy of the probabilistic models estimation. By comparing with some state-of-the-art content-based image labeling methods, we have empirically shown that 1) the proposed method can achieve comparable tag prediction accuracy when there is no user-provided tag, and that 2) it can significantly boost the prediction accuracy if the user can provide just a few tags.

1. Introduction

Photo sharing on the Internet has become very popular and the numbers of photos being uploaded onto these sites are increasing at a rapid speed. How to manage such huge collections of photos to enable users to find their interested photos quickly is a very challenging task. In fact, in the past decade, image retrieval and managing large image repositories have attracted extensively research interest across many disciplines in computer science including artificial intelligence, computer vision, database, etc. One of the approaches that has been intensively researched is content-based image retrieval (CBIR) [14] where image retrieval is performed based on the visual similarity of low-level image features such as color, textures, object shapes, etc. However, the practical success of CBIR has been pretty limited due to the inconsistency between low-level visual similarity and high-level perceived subjective image similarity which is often referred to as the semantic gap [14]. One way to reduce the semantic gap is by introducing high-level knowledge through user labeling or tagging the images. Many multimedia content sharing platforms, e.g., Flickr [5], PhotoStuff [6], provide annotation functions for their users to tag the images manually. While the carefully provided tags can enable accurate semantic image retrieval, the tagging process is tedious and labor-intensive [10].

What is desired is a method which can automatically or semi-automatically label the images. Recently, methods aim to automatically produce a set of semantic tags for images based on their visual contents have attracted a lot of attention [2], [3], [4], [8], [9]. These methods first extract low-level features of the images and then build a mathematical model to associate these low-level image contents with tags. We refer to such methods as *content-based image labeling*.

Going beyond the content-based approach, collaborative filtering is an alternative that explores the correlation between user related attributes (e.g., ratings given by the users, usage patterns, etc.) of various information items for filtering or recommendation. It does not require directly looking into the actual contents of the information items [7]. Recently, the collaborative approach has already been applied to image retrieval successfully [15]. In particular, given a few tags provided by the users, additional tags of the image can be predicted by leveraging the tag-to-tag correlation. For instance, with two groups of images - one tagged with "sky" and "tree" and another tagged with "tree" and "grass", a new image tagged with only "grass" will be predicted to have the tag "sky" even though "grass" and "sky" have never been tagged to the same image by the users. This collaborative approach, however, requires the existence of a collection of carefully prepared user-provided tags which could be lacking initially in many cases. This is the so-called cold start problem. To alleviate this, methods that combined content-based and collaborative filtering have been proposed (e.g., [13]).

In this paper, we present a novel hybrid approach to automatic image labeling which integrates low-

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

level image visual content information and high-level user-provided tags so as to take advantages of both content-based and collaborative approaches. A unified probabilistic framework is derived for the integration. In particular, the image visual contents are represented using a colored pattern appearance model (CPAM) [12] and visual feature probability distributions of different semantic concepts are learned via nonparametric density estimation with kernel smoothing. User-provided tags are incorporated into this framework by estimating tag co-occurrence probabilities using collaborative filtering. We evaluated the proposed approach based on a widely used Corel data set. By comparing with some state-of-the-art content-based image methods, we have empirically shown that 1) the proposed method can achieve comparable tag prediction accuracy when there is no user-provided tag, and that 2) it can significantly boost the prediction accuracy if the user can provide just a few tags.

2. Collaborative and Content-based Integration - A Probabilistic Approach

In order to integrate visual contents and user-provided tags, we propose a probabilistic framework to support automatic image labeling.

2.1. The Framework

Let $\mathcal{W} = \{w_1, w_2, \ldots, w_M\}$ be the word vocabulary of the labeling system and $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ be the global feature vector representing the visual content of an image computed based on CPAM [12]. Also, assume that each image I_i are associated with T user-provided tags[1], denoted as $\mathcal{W}_i^u = \{w_{i_1}, w_{i_2}, \ldots, w_{i_T}\}$, with $\mathcal{W}_i^u \subset \mathcal{W}$. Automatic labeling can thus be formulated as selecting from $\mathcal{W}_i^c = \mathcal{W} \backslash \mathcal{W}_i^u$ the words which give high values of $P(w_j|I_i, \mathcal{W}_i^u)$ as the tags of the image I_i. To compute $P(w_j|I_i, \mathcal{W}_i^u)$, i.e., the probability that I_i is tagged with w_j,

$$\begin{aligned} P(w_j|I_i, \mathcal{W}_i^u) &= P(w_j|\mathbf{x}, \mathcal{W}_i^u) \quad (1) \\ &= \frac{P(\mathbf{x}|w_j)P(\mathcal{W}_i^u|w_j, \mathbf{x})P(w_j)}{P(\mathbf{x})P(\mathcal{W}_i^u|\mathbf{x})}. \end{aligned}$$

By assuming that the occurrence of user-provided tags is independent of particular images, Eq. 1 is rewritten as

$$P(w_j|\mathbf{x}, \mathcal{W}_i^u) = \frac{P(\mathbf{x}|w_j)P(\mathcal{W}_i^u|w_j)P(w_j)}{P(\mathbf{x})P(\mathcal{W}_i^u)}. \quad (2)$$

1. If an image is not labeled with any tag initially, i.e. there is no user-provided tag available, $T = 0$.

Also, by assuming that the tags in $\mathcal{W}_i^u$ are mutually independent given any w_j, $P(\mathcal{W}_i^u|w_j)$ can be rewritten as

$$P(\mathcal{W}_i^u|w_j) = \prod_{t=1}^{T} P(w_{i_t}|w_j). \quad (3)$$

Taking the logarithm of Eq. 2, we have

$$\begin{aligned} &\log P(w_j|\mathbf{x}, \mathcal{W}_i^u) \\ &= \log\left(\frac{P(\mathbf{x}|w_j)\prod_{t=1}^{T} P(w_{i_t}|w_j)P(w_j)}{P(\mathbf{x})P(\mathcal{W}_i^u)}\right) \\ &= \log P(\mathbf{x}|w_j) + \sum_{t=1}^{T} \log P(w_{i_t}|w_j) \\ &\quad + \log P(w_j) - \log P(\mathbf{x}) - \log P(\mathcal{W}_i^u). \quad (4) \end{aligned}$$

By computing the probability for each word $w_j \in \mathcal{W}_i^c$ according to Eq. 4, all the candidate words can be ranked as suggestions to label the image. In Eq. 4, $P(\mathbf{x})$ and $P(\mathcal{W}_i^u)$ are constants, and thus can be ignored. $P(w_j)$ is the prior probability of word w_j which can be easily estimated using the training set or kept uniform. What we have to estimate are $P(\mathbf{x}|w_j)$ and $P(w_i|w_j)$.

2.2. Nonparametric Density Estimation for $P(\mathbf{x}|w)$

We interpret $P(\mathbf{x}|w)$ as the distribution of visual feature $\mathbf{x}$ conditional upon the assignment of semantic concept w. To estimate the density, a non-parametric density estimation approach with kernel smoothing [11] is adopted. Assuming $\mathcal{D}_w = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ to be the set of training samples extracted from the images with the label w, we estimate

$$P(\mathbf{x}|w) = \frac{1}{n}\sum_{i=1}^{n} k\left(\mathbf{x} - \mathbf{x}^{(i)}; h\right), \quad (5)$$

where k is a D-dimensional Gaussian kernel that we place on each point, given as

$$k(\mathbf{t}; h) = \prod_{d=1}^{D} \frac{1}{h_d\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\mathbf{t}_d}{h_d}\right)^2\right) \quad (6)$$

and, the bandwidth parameters $\{h_d\}$ are selected empirically on a held-out set of the training set.

2.3. Estimating $P(w_j|w_i)$ Using A Collaborative Approach

To estimate $P(w_j|w_i)$, a probability table based on tag co-occurrence is first computed [8]. Assuming that there are N tagged images with M distinct tags in the

training set, the image-tag pairs can be represented as an N–by–M matrix U, where each row corresponds to a particular image and each column corresponds to a particular tag. The element $U(i,j)$ is 1 if the i^{th} image is labeled with the j^{th} tag, and 0 otherwise. As it is unusual for a user to provide a large number of tags for each training image, the image-tag matrix U is usually very sparse. Using it to directly estimate $P(w_j|w_i)$ will give inaccurate results. With the conjecture that images with a number of tags in common are highly likely to be contextually similar, the set of tags associated with these contextually similar images could be "shared" among them. The idea is essentially the same as the underlying principle of collaborative filtering [1]. Thus, $P(w_j|w_i)$ is estimated as depicted in Algorithm 1 and each element $T_{corr}(i,j)$ can then be used as the estimate of $P(w_j|w_i)$.

Algorithm 1 $P(w_j|w_i)$ estimation using a collaborative approach.

Input: The original image-tag matrix U
Output: Probability table T_{corr}

1: Apply the collaborative filtering method similar to [16] on U and obtain U_*.
2: Compute $U_*^T \times U_*$ that gives a M–by–M matrix and normalize each row to get T_{corr}.

3. Experiments

To evaluate the performance of the proposed hybrid approach, experiments have been conducted using a popular Corel data set [3] which contains 5,000 images (4,500 for training and 500 for testing) and has been used in [2], [3], [4], [9] for benchmarking. Each image is tagged with 1-5 words and there are altogether 371 distinct words in the vocabulary. The mean number of words per image is 3.5, which reflects that the image-tag matrix is indeed sparse.

To demonstrate how the proposed model can effectively exploit a small number of user provided tags to greatly enhance the tag prediction accuracy, we randomly selected T (=0, 1, 2) tags for each test image as the user-provided tags and then attempted to predict the remaining tags. Following the conventions commonly used in the collaborative filtering literature, we term these protocols $Given\ T$ [1].

Similar to [2], [3], [4], [9], we computed the *five* words which give the largest values based on Eq. 4 under the $Given\ 0$ protocol. In $Given\ 1$ and $Given\ 2$ protocols, the annotation length is set to be *four* and *three*, respectively. We then computed the recall and precision rates per word for the test set. In particular, for a given tag w, let N_h^w be the number of images in the test set that have been labeled with the tag, N_{sys}^w be the number of images that are predicted to be associated with the tag by our system, and N_c^w be the number of images correctly tagged by our system. The precision and recall rates for w are defined as $recall(w) = \frac{N_c^w}{N_h^w}$ and $precision(w) = \frac{N_c^w}{N_{sys}^w}$, respectively. We present the average recall and precision rates over all words in the test set.

Table 1. Performance comparison between the cases with and without collaborative filtering (CF) employed for the estimation of the co-occurrence probabilities of word pairs.

Protocols	$Given\ 1$		$Given\ 2$	
Hybrid Model	Without CF	With CF	Without CF	With CF
Avg. Recall.	0.24	0.28	0.30	0.33
Avg. Prec.	0.16	0.18	0.19	0.21

In Table 1, we tabulated the performance of the proposed approach that estimates the word co-occurrence probability $P(w_j|w_i)$ with and without collaborative filtering (CF). It is clearly shown that the use of the CF method introduced in Section 2.3 is effective in alleviating the tag sparsity problem and can boost the performance significantly.

In Table 2, we present the performance comparison between the proposed approach and those previously proposed in the literature that used the same data set for evaluation. They include the translation model [3], the continuous-space relevance model (CRM) [9], the multiple-Bernoulli relevance model (MBRM) [4], and the supervised multiclass labeling (SML) [2]. Under the $Given\ 0$ protocol, only low-level visual features (CPAM-based) are used in our approach (denoted as HM, $Given\ 0$) and the performance was found to be comparable to that of CRM. However, given only one tag (the $Given\ 1$ protocol), the performance of the proposed approach was greatly improved and found to be very close to that of SML. Under the $Given\ 2$ protocol, our proposed approach achieves the best recall rate and its precision rate is only slightly lower than those of MBRM and SML.

4. Conclusions and Future Work

In this paper, we derived a probabilistic framework for fusing low-level image visual contents and high-level user-provided tags to perform automatic image labeling. We have shown that collaborative filtering techniques can effectively alleviate the tag sparsity problem for better estimation of tag co-occurrence probability as required in the framework. By evaluating

Table 2. Performance comparison with some representative image labeling methods based on the same Corel data set [3].

Models	Translation	CRM	MBRM	SML	HM, $Given\ 0$	HM, $Given\ 1$	HM, $Given\ 2$
Avg. Recall.	0.04	0.19	0.25	0.29	0.14	0.28	0.33
Avg. Prec.	0.06	0.16	0.24	0.23	0.10	0.18	0.21

the proposed approach based on a benchmark data set, we demonstrated that by requiring the user to provide only a few tags, drastic improvement in tag prediction over purely content-based methods can be achieved. Our future work includes enhancing the computational efficiency of the proposed approach for large-scale photo repositories such as Flickr.

5. Acknowledgment

The authors would like to thank Kobus Barnard for providing the Corel data set [3]. This work was supported in part by MoE research Fund under contract 104075, Shanghai Municipal R&D Foundation under contract 06DZ15008, and MoST Support Program under contract 2007BAH09B03. This research was performed at Hong Kong Baptist University.

References

[1] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI'98)*, pages 43–52, 1998.

[2] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans Pattern Anal Mach Intell*, 29(3):394–410, March 2007.

[3] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object recognition as machine translation :learning a lexicon for a fixed image vocabulary. In *Proc. of the 7th European Conference on Computer Vision (ECCV'02)*, 2002.

[4] S. Feng, V. Lavrenko, and R. Manmatha. Multiple bernoulli relevance models for image and video annotation. In *Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR'04)*, volume 2, pages 1002–1009, 2004.

[5] Flickr. http://www.flickr.com, Yahoo!, 2005.

[6] C. Halaschek-Wiener, J. Golbeck, A. Schain, M. Grove, B. Parsia, and J. Hendler. Photostuff-an image annotation tool for the semantic web. In *Proc. of 4th Intl. Semantic Web Conference*, 2005.

[7] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *Proc. of the 25th Annual Intl. ACM SIGIR Conf. (SIGIR'99)*, pages 230–237, New York, NY, USA, 1999. ACM Press.

[8] Y. Jin, L. Khan, L. Wang, and M. Awad. Image annotations by combining multiple evidence & wordnet. In *Proc. of the 13th annual ACM international conference on Multimedia*, Hilton, Singapore Nov. 06–11, 2005.

[9] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Proc. of Advances in Neutral Information Processing Systems (NIPS'03)*, 2003.

[10] W. Liu., S. Dumais, Y. Sun, H. Zhang, M. Czerwinski, and B. Field. Semi-automatic image annotation. In *INTERACT2001, 8th IFIP TC.13 Conference on Human-Computer Interaction*, Tokyo, Japan Jul. 9–13, 2001.

[11] E. Parzen. On estimation of a probability density and mode. *Annals of Mathematical Statistics*, 35:1065–1076.

[12] G. Qiu. Indexing chromatic and achromatic patterns for content-based colour image retrieval. *Pattern Recognition*, 35:1675–1685, August 2002.

[13] A. Schein, A. Popescul, L. Ungar, and D. Pennock. Methods and metrics for cold-start recommendations. In *Proc. of the 25th Annual Intl. ACM SIGIR Conf. (SIGIR'02)*, pages 253–260, 2002.

[14] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1349–1380, 2000.

[15] S. Uchihashi and T. Kanade. Content-free image retrieval by combinations of keywords and user feedbacks. In *Proc. of the 4th Intl. Conf. on Image and Video Retrieval*, pages 650–659, 2005.

[16] S. M. Weiss and N. Indurkhya. Lightweight collaborative filtering method for binary-encoded data. In *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2001)*, pages 484–491, Sep. 03–05, 2001.