ANNOTATING IMAGES BY HARNESSING WORLDWIDE USER-TAGGED PHOTOS

Xirong Li, Cees G.M. Snoek, and Marcel Worring

ISLA, Informatics Institute, University of Amsterdam
Science Park 107, 1098 XG, Amsterdam, The Netherlands
{x.li,cgmsnoek,m.worring}@uva.nl

ABSTRACT

Automatic image tagging is important yet challenging due to the semantic gap and the lack of learning examples to model a tag's visual diversity. Meanwhile, social user tagging is creating rich multimedia content on the web. In this paper, we propose to combine the two tagging approaches in a search-based framework. For an unlabeled image, we first retrieve its visual neighbors from a large user-tagged image database. We then select relevant tags from the result images to annotate the unlabeled image. To tackle the unreliability and sparsity of user tagging, we introduce a joint-modality tag relevance estimation method which efficiently addresses both textual and visual clues. Experiments on 1.5 million Flickr photos and 10 000 Corel images verify the proposed method.

Index Terms— Automatic image tagging, User tagging

1. INTRODUCTION

The advent of user-generated visual content, e.g., homemade photos and video clips shared on the web, is demanding effective and scalable solutions to handle such increasing amounts of diverse multimedia data. To meet the demand, on one hand, much effort is devoted to making computers "understand" the content. As an important instance towards this direction, *machine tagging* targets at annotating images by computer. Nonetheless, machine tagging is still very challenging, due to the semantic gap [8]. Images of the same concept vary significantly in terms of visual appearance, e.g., illumination, scale, and perspective. A large and diverse set of learning examples is imperative to model the visual diversity. On the other hand, social multimedia sharing systems have successfully motivated common users around the world to tag their visual content on the web. A good example of *user tagging* is Flickr. It hosts over 2 billion images, and receives around 3 million new uploaded photos per day. However, tags contributed by users are known to be ambiguous, limited in terms of completeness, and overly personalized [4]. This is not surprising due to the significant diversity of users' knowledge and cultural background. Given web-scale, yet unreliable, user-tagged photos, an interesting question is whether we can harness user tagging to solve the image tagging problem.

Many methods have been proposed to tackle the image tagging problem. We divide them according to their model-dependence into two types of approaches, namely model-based approaches and model-free approaches. Given a set of labeled images as training data, the model-based approaches focus on learning a mapping between low-level visual features (e.g., color and local descriptors) and high-level semantic concepts (e.g., airplane and classroom), e.g., [1, 2, 5]. Due to the expense of manual labeling, however, currently only a limited number of visual concepts can be modeled effectively using small-scale datasets. Besides, the approaches are often computationally expensive, making them difficult to scale up. We refer to [3] for more discussions about the model-based approaches. In contrast, the second type of approaches attempts to annotate an image in a model-free way by utilizing web photos, e.g., [9–11]. The approaches assume there exists a large well-labeled database such that one can find a visual duplicate for the unlabeled example. Then, automatic tagging is done by simply propagating tags from the duplicate to that image. However, since the database in reality is of limited-scale with noisy annotation, neighbor search is first conducted to find visual neighbors. De-noising methods are then used to select relevant tags, out of raw annotations of the neighbors, to annotate the unlabeled image.

Inspired by the initial success of the model-free methods in handling web-scale data, we propose to combine machine tagging and user tagging by employing the model-free methodology. As a test case, we choose Flickr as an example of user-tagging and a search-based framework by Wang et al. [10] as an instance of machine tagging. The novelty of this work is that we introduce into the search-based framework an effective joint-modality tag relevance estimation method. By taking both textual and visual clues into account, the method accurately estimates tag relevance from 1.5 million user-tagged images.

The rest of the paper is organized as follows. We describe the proposed method in Section 2, followed by experiments in Section 3. We conclude the paper in Section 4.

This work was supported by the EC-FP6 VIDI-Video project and the STW SEARCHER project.

2. MACHINE-AND-USER TAGGING

2.1. Search-based Machine Tagging

We aim for machine tagging methods that accurately predict relevant tags with respect to unlabeled images. We consider a tag relevant to an image if the tag accurately describes objective aspects of the visual content, or in other words, users with common knowledge relate the tag to the visual content easily and consistently. Although the relevance of a tag given the visual content can be subjective for a specific user, an objective criterion is desirable for the general content understanding problem. Given an unlabeled image I_q, we seek a set of keywords $\mathbf{w}^*$ most relevant with respect to I_q, i.e.,

$$\mathbf{w}^* = \arg\max_{\mathbf{w}\subset\mathcal{V}} rel(\mathbf{w}|I_q),$$

where $rel(\mathbf{w}|I_q)$ is a measurement of tag relevance and $\mathbf{w}$ keywords in a predefined vocabulary $\mathcal{V}$.

If a well-labeled duplicate of image I_q exists, machine tagging is solved by using tags of the duplicate. In reality, however, the assumption is often violated: either the duplicate does not exist, or it is unlabeled or mislabeled. An intuitive idea is then to approximate the duplicate by a set of visually similar examples. Despite the semantic gap, Torralba et al. [9] found out that given a very large dataset (80 million tiny images in their experiments), one might find relevant images in visual neighbors with a reasonable accuracy. Meanwhile, the increasing amounts of web photos tagged by users around the world is creating a potentially unlimited-scale database. Though visual search results can be unsatisfactory, if relevant tags stand out from relevant images, one might still find prospective tags for tagging. However, user tagging is known to be subjective and unreliable. Moreover, individual tags are mostly used once per image by user tagging. This implies that within an image, relevant tags and irrelevant ones are not distinguishable by their occurrence frequency. Identifying relevant tags from neighbor images is thus crucial.

In this work, we adopt a search-based approach by Wang et al. [10] to harness web-scale user-tagged photos for machine tagging. Let d be a visual distance function between two images. For image I, we denote its k nearest neighbors found in a user-tagged database in terms of d as $NN_d(I,k)$. In the search-based approach,

$$\begin{aligned}\mathbf{w}^* &\approx \arg\max_{\mathbf{w}\subset\mathcal{V}} rel(\mathbf{w}|NN_d(I_q,k)) \qquad (1)\\ &\approx \arg\max_{\mathbf{w}\subset\mathcal{V}} \sum_{J\in NN_d(I_q,k)} rel(\mathbf{w}|J)\cdot sim(J,I_q),\end{aligned}$$

where $sim(J,I_q)$ is a measurement of semantic similarity between J and I_q. In particular, we annotate I_q by a two-step procedure as illustrated in Fig. 1, that is,

• *Step 1. Search by content*. We find k nearest neighbors of I_q from a user-tagged image database.

• *Step 2. Tag relevance estimation*. Given tags of the neighbor images, we then select most relevant tags to annotate I_q.

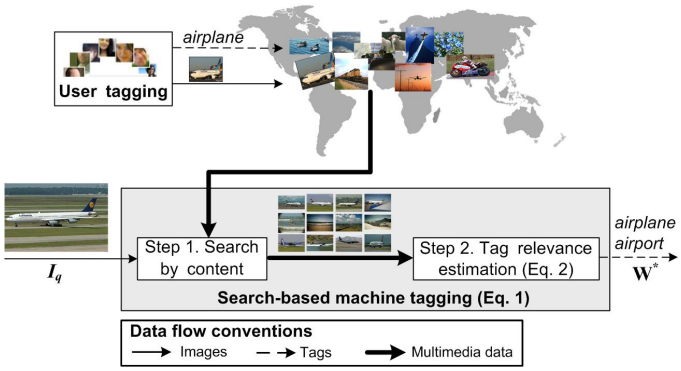

Fig. 1. **Annotating images by harnessing worldwide user-tagged photos within a search-based framework**. The main contribution of this work is the proposed joint-modality tag relevance estimation method that simultaneously addresses textual and visual clues in a scalable way.

2.2. Tag Relevance Estimation from User Tagging

In previous work, methods to estimate the relevance of a tag given an image are mainly derived from the text retrieval field. In [10,11], for instance, term frequency-inverse document frequency (tf-idf) is adapted to rank and select relevant tags. In the tf-idf weighting scheme, the relevance of tag w with respect to image J is calculated as

$$rel(w|J) = tf(w,J)\cdot idf(w),$$

where $tf(w,J)$ is occurrence frequency of w in tags of J. The function $idf(w)$ is calculated as $\frac{1}{\log_2(df(w)+1)}$, where $df(w)$ is the number of images labeled with w in the entire collection. Note that tf and idf measure the tag's importance within an image and its informativeness within the collection, respectively. Despite the success of the tf-idf principle in text retrieval, the unreliability and sparsity of user tagging make the text-based methods inaccurate.

Intuitively, if one can effectively exploit both textual and visual information, tag relevance estimation could be more accurate. However, visual features are often of high dimensionality and significant diversity exists in user tagging vocabulary. Hence, directly modeling co-occurrence of textual and visual modalities, e.g., using a multivariate Gaussian, tends to be problematic. As an alternative, we introduce a non-parametric joint-modality method based on a neighbor voting algorithm proposed in our earlier work [6]. The algorithm originally aims for social image retrieval. In this paper, we demonstrate the general applicability of the algorithm by extending it to the machine tagging scenario. The intuition behind neighbor voting is, if different persons label visually similar images using the same tags, these tags are likely to reflect objective aspects of the visual content. Hence, the relevance of a tag with respect to an image can be inferred from tagging behavior of visual neighbors of that image. Given

image I labeled with tag w, we show in [6] that estimating the relevance of w with respect to I amounts to counting the number of w in visual neighbors of I, i.e.,

$$rel(w|I) = |\{J \in \mathcal{I} | J \in NN_d(I, n), w \in \mathbf{w}_J\}|, \quad (2)$$

where n is the number of neighbors used for voting, and $|\bullet|$ the cardinality operator on image sets. Through the algorithm, unambiguous and objective tags receiving most neighbor voting stand out. We further multiply the relevance value by $idf(w)$ to take tag informativeness into account. Since the text-based methods ignore visual clues, they are conducted in a global visual feature space. In contrast, our method computes tag relevance in a localized space, i.e., the neighborhood of image I. It is this restriction that provides a joint-modality mechanism, leading to more accurate and robust tag relevance estimation.

Finally, to annotate an image, we choose all tags from its neighbors to form a candidate tag set. Each tag in the set is ranked in descending order according to its relevance value computed using Eq. 1. We approximate $sim(J, I_q)$ by using visual dissimilarity , i.e., $sim(J, I_q) = e^{\frac{-d(J,I_q)^2}{2}}$ as suggested in [10]. We select t top ranked tags as the final annotation. The choice of t is a tradeoff between annotation precision and recall, which needs to be determined experimentally.

3. EXPERIMENTS

3.1. Experimental Setup

User-tagged image database. We downloaded 1.5 million tagged images from Flickr using its API service (`http://www.flickr.com/services/api/`). The images are of medium size with maximum width or height fixed to 500 pixels. The number of distinct tags per image varies from 2 to 1231, with an average value of 8. By removing rare tags that are used less than 5 times in the entire collection, we get 90 346 unique tags.

Evaluation set. We use 10 000 well-labeled Corel images as an evaluation set. The number of distinct tags per image varies from 1 to 14, with an average value of 6. There are 5,469 unique tags in total (after Porter stemming). Since Corel images are professional stock photos while Flickr images are mostly personal pictures taken by common users, the two datasets are different in terms of visual similarity. Hence, our experimental setting is much closer to a real scenario and thus more challenging than a popular yet heavily criticized setting using 90% Corel images for training and remaining 10% images for evaluation (see Müller et al. [7]).

Evaluation criteria. We employ two standard criteria to evaluate the annotation performance, i.e., precision and recall. Given a test image I,

$$\begin{aligned} \text{precision}(I) &= \frac{\text{Number of correctly predicted tags}}{\text{Number of predicted tags}}, \\ \text{recall}(I) &= \frac{\text{Number of correctly predicted tags}}{\text{Number of ground-truth tags}}. \end{aligned}$$

For a test set consisting of n images, precision and recall are averaged over all test images. To study the coverage ability of an automated image tagging method, we further define

$$\text{coverage} = |\{I \in \text{test set} | \text{precision}(I) > 0\}|/n.$$

Visual feature. Since we need features relatively stable for search and efficient to compute to handle millions of images, we use a 64-dimensional global color-texture feature [6]. The dissimilarity between images is measured using the Euclidean distance between feature vectors. In order to retrieve visually similar images from the 1.5M image set efficiently, we employ a parallel K-mean clustering strategy for speed-up [6].

We conduct two experiments for evaluation.

Experiment-1: Joint-modality versus Text. We compare our joint-modality method with a text-based baseline [10]. For each method, we study its performance with the number of neighbor images k being 20 and 200, respectively. We set t, i.e., the number of predicted tags per test image, to 500 and calculate interpolated recall-precision curves for the test set. The number of neighbors for the neighbor voting algorithm is fixed to 1000 throughout the experiments.

Experiment-2: The impact of user-tagged database size. Since the amount of user-tagged images increases rapidly, we investigate whether the annotation will improve as the database grows. We conduct a simulated experiment by setting the size of the Flickr database to 15K, 150K, and 1.5M.

3.2. Results

Experiment-1: Joint-modality versus Text. The joint-modality method consistently outperforms the text-based method, given the same number of neighbor images (see Fig. 2). When the neighbor set is small, the former is significantly better than the latter. Since relevant images are rare in a small neighbor set, this result confirms our intuition that if relevant tags stand out from relevant images, good tagging quality is still expectable even the visual search is unsatisfactory. Besides, the joint-modality method using 20 neighbors is comparable to the text-based method using 200

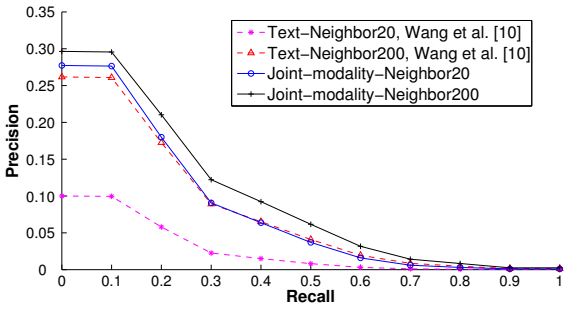

Fig. 2. **Experiment-1: Joint-modality versus Text**. Comparison between different tag relevance estimation methods.

Table 1. Experiment-1: Joint-modality versus Text. For each test image, we select the 5 top ranked tags as the final annotation. The number of neighbor images k is fixed to 200. *Std.* is the standard deviation of *Precision* and *Recall*.

	Evaluation criteria				
Method	*Precision*	*Std.*	*Recall*	*Std.*	*Coverage*
Text [10]	0.093	0.144	0.081	0.128	0.351
Joint-modality	0.111	0.153	0.096	0.136	0.411

neighbors. Note that the off-line tag relevance learning has to be performed only once. While the on-line search time is at least linear to the number of neighbors required. Hence, for on-line applications our joint-modality method is more efficient, which is a big advantage for real-time scenarios. We further show in Table 1 the effectiveness of the proposed method by fixing the number of predicted tags to 5.

Experiment-2: The impact of user-tagged database size. The performance improves as we increase the number of Flickr images, as shown in Fig. 3. Given a query example, one might find more relevant images from a larger database, as observed in [9]. Besides, a larger database also improves estimation accuracy of the joint-modality method. We observe that there is a significant improvement when we scale up the database from 15K to 150K, while the improvement from 150K to 1.5M is relatively small. An open question remains, i.e., how many images are needed to make the visual feature space sufficiently dense such that an image will find relevant ones within its neighbors? We finally present several machine tagging examples in Fig. 4.

4. CONCLUSIONS

In this paper, we tackle the problem of automatic image tagging by harnessing worldwide user-tagged images. In particular, we adopt a search-based approach to combine machine tagging and user tagging. To cope with subjective user tagging, we introduce a joint-modality tag relevance estimation method that simultaneously takes textual and visual clues into account. By identifying and reinforcing relevant tags from search results, the method makes the search-based approach more robust then a text-based alternative to the semantic gap problem. Experiments on 1.5 million Flickr photos and 10 000 Corel images demonstrate the viability of our approach. Still, the semantic gap problem needs to be addressed under the current framework. We will investigate the potential for improving the consistency between visual similarity and semantic similarity by including more advanced visual features.

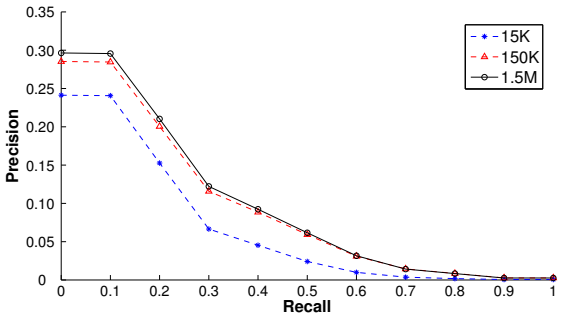

Fig. 3. Experiment-2: The impact of user-tagged database size. For each test image, we find 200 nearest neighbors and use the joint-modality method to select 500 relevant tags.

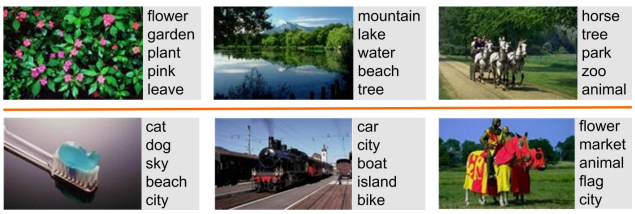

Fig. 4. Images tagged automatically by our method. Good examples are at the top row and bad ones at the bottom row.

5. REFERENCES

[1] K. Barnard, P. Duygulu, D. Forsyth, N. Freitas, D. Blei, and M. Jordan. Matching words and pictures. *JMLR*, 3(6):1107–1135, 2003.

[2] E. Chang, G. Kingshy, G. Sychay, and G. Wu. CBSA: content-based soft annotation for multimodal image retrieval using bayes point machines. *TCSVT*, 13(2):26–38, 2003.

[3] R. Datta, D. Joshi, J. Li, and J. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Survey*, 40(2):1–60, 2008.

[4] S. Golder and B. Huberman. Usage patterns of collaborative tagging systems. *Information Science*, 32(2):198–208, 2006.

[5] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *TPAMI*, 30(6):985–1002, 2008.

[6] X. Li, C. G. M. Snoek, and M. Worring. Learning tag relevance by neighbor voting for social image retrieval. In *ACM MIR*, pages 180–187, 2008.

[7] H. Müller, S. Marchand-Maillet, and T. Pun. The truth about Corel - evaluation in image retrieval. In *CIVR*, pages 38–49, 2002.

[8] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *TPAMI*, 22(12):1349–1380, 2000.

[9] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *TPAMI*, 30(11):1958–1970, 2008.

[10] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Scalable search-based image annotation. *Multimedia Systems*, 14(4):205–220, 2008.

[11] X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma. Annotating images by mining image search results. *TPAMI*, 30(11):1919–1932, 2008.