SEARCH-BASED AUTOMATIC IMAGE ANNOTATION VIA FLICKR PHOTOS USING TAG EXPANSION

Liang-Chi Hsieh, Winston H. Hsu

Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan

ABSTRACT

Exponentially growing photo collections motivate the needs for automatic image annotation for effective manipulations (e.g., search, browsing). Most of the prior works rely on supervised learning approaches and are not practical due to poor performance, out-of-vocabulary problem, and being time-consuming in acquiring training data and learning. In this work, we argue automatic image annotation by search over user-contributed photo sites (e.g., Flickr), which have accumulated rich human knowledge and billions of photos. The intuition is to leverage surrounding tags from those visually similar Flickr photos for the unlabeled image. However, the tags are generally few and noisy. To tackle such challenges, we propose a novel solution in three folds: (1) a tag expansion method to solve the sparsity of user-contributed tags; (2) improving tag relevance estimation by visual consistency between candidate annotations and the unlabeled image, and (3) the semantic tag consistence among candidate tags. Experimenting over Flickr photo benchmarks and requiring no additional keywords, we show that the proposed method significantly outperforms prior works and even provide more diverse annotations.

Index Terms— Search-based automatic image annotation, tag expansion

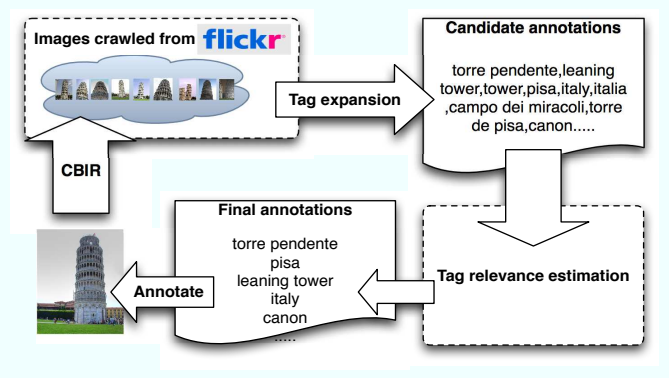

Fig. 1. The proposed automatic image annotation process. For a unlabeled image, visually similar photos in social media (i.e., Flickr) are first retrieved by a CBIR system. The associated tags from these images are then expanded (by Google snippets) as candidate annotations, which are later ranked by estimating their relevances to the given image in terms of visual consistency and semantic consistency.

1. INTRODUCTION

There arise the needs for effective manipulations (e.g., search, browsing, etc.) for exponentially growing photo collections with the prevalence of photographing devices and the popularity of media-sharing services. *Image annotation* – giving semantically relevant words to an image – is one of the enabling technologies for bridging the semantic gap in prior applications. However, the tedious and time-consuming processes for manual annotation motivate the crucial researches in automatic image annotation.

Some previous works [1, 2] in (supervised) image categorization and classification are first proposed to annotate images. However, due to the fixed and limited vocabularies in pre-defined ontology, it is impractical to deploy for consumer photos; meanwhile, such supervised learning methods require huge human-labelled data and are difficult to scale.

Recently, researchers [3, 4, 5] begin to leverage the rich resources on the Web for automatic image annotation. Such methods adopt annotations from the surrounding texts of visually similar images crawled from the Web and usher a promising solution to automatic image annotation with unlimited and up-to-date vocabularies. The similar photos are retrieved and ranked by content-based image retrieval (CBIR) systems. The Web descriptions are generally rich with hundreds of words. AnnoSearch proposed in [3] annotates an input image associated with an initial keyword by leveraging the surrounding texts of images crawled from the Web. Search result clustering is applied on the surrounding texts to obtain annotations from the frequent words in each cluster. The major assumption for AnnoSearch is an initial and accurate keyword associated with the input image. However, such keyword is not always available. Wang et al. remove such assumption in the derived work [4] and introduce two strategies for calculating the relevance score of an annotation by (modified) frequency counts as well.

The emerging social media (e.g., Flickr, YouTube, etc.) accommodate enormous photos and videos augmented with rich context such as user-provided tags, geo-locations, time, device metadata, etc., and have been shown benefiting a wide variety of potential applications such as annotation, recommendation, and retrieval [6, 7, 8]. Li et al. [6] try to use photos and tags in Flickr to automatically annotate images. They propose a tag ranking method by counting tag occurrences in the visually similar photos. Though promising, they ignore the noise and low accuracy problems commonly observed in user-generated tags in social media [7] (also cf. Figures 2 and 3). In contrast to Web images, Flickr photos are usually associated with sparse text tokens (e.g., averagely fewer than 3 tags as shown in [7]). Wu et al. [9] propose a learning-based tag recommendation scheme for Flickr photos leveraging tag co-occurrence and visual-based tag correlation. However, their approach still needs initial keywords associated with input image. It finds related tags based on existing keywords of image, but can not be applied on image without any keywords.

Leveraging accumulated knowledge in social media, we argue to tackle the above problems by proposing a search-based automatic image annotation as illustrated in Figure 1. In order to deal with

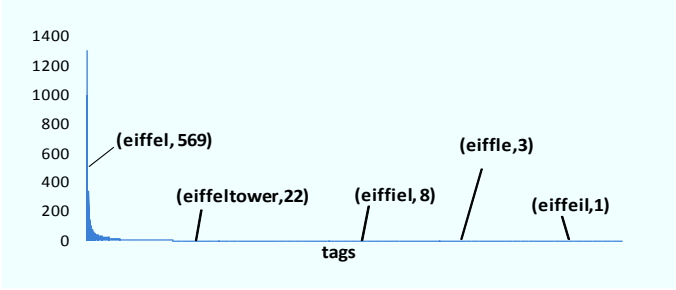

Fig. 2. Tag frequency distribution in our sampled Flickr image dataset. For describing the same semantic target, very diverse tags are usually observed due to user-contributed annotations. For example, in Eiffel-related terms, we observe that several different spellings, followed by their frequency counts, are shown in the figure.

the sparsity and unreliability of user-provided tags, we emphasize the novel solution in three folds: (1) a tag expansion method using Google snippets to expand sparse user tags on Flickr photos; (2) improving tag relevance estimation by considering visual consistency between candidate annotations and the unlabeled image, and (3) the semantic consistency among the candidate tags. Based on them, given the un-annotated image, we can collect more semantically relevant tags by expansion from those limited (or noisy) tags along with the visually similar Flickr photos; such candidate tags can then be ranked by evaluating visual and semantic consistencies. Experimenting in a large Flickr dataset, we show that the proposed method significantly outperforms the prior methods in different performance metrics.

2. THE PROPOSED METHOD

The goal of automatic image annotation is to accurately predict relevant annotations with respect to an unlabeled image. Provided the image, we can only rely on the visual content to predict annotations. Given an unlabeled image I_q, a set of keywords w^* is selected from the candidate annotation set T if they are most relevant to I_q, i.e.,

$$w^* = \underset{w \in T}{\operatorname{argmax}}\, rel(w|I_q), \tag{1}$$

where $rel(w|I_q)$ is the measurement of tag relevance in semantics with I_q. The intuition for search-based image annotation is that we can use tags collected from visually similar Flickr photos (i.e., T) to approximately annotate the image I_q. In the case, the most important problem is to find most relevant tags from these similar Flickr photos, i.e. an auxiliary knowledge source for annotation. However, the noisy tags extracted from these photos make finding relevant tags nontrivial. Moreover, the automatic image annotation suffers from the sparsity of user-generated tags. Recognizing these problems, we propose to annotate an unlabeled input image I_q by a two-step procedure:

Tag expansion: We use the search results from Google to expand sparse user tags of visually similar consumer photos; i.e., promising to improve the annotation recall rate by increasing the candidate set T (See Section 2.1).

Candidate annotation ranking: We estimate tag relevance with respect to image by visual and semantic consistency and use the relevance scores to rank tags. Here we approximate $rel(w|I_q)$ in Eq. 1 by tag relevance score $rel(I_q, w)$ explained in Section 2.2.

1) Tour-eiffel, 0.67	6) Tour-de-eiffel, 0.58	11)Eiffellight, 0.45
2) Torre-eiffel, 0.66	7) Visittoureiffel, 0.50	12) Eiffle, 0.41
3) Toureiffel, 0.65	8) Visteiffel, 0.50	13) Effiel, 0.41
4) Eiffeltower, 0.63	9) Sparklyeiffel, 0.50	14) Eiffeil, 0.40
5) Eiffel-tower, 0.62	10) Eiffiel, 0.46	15) Effel, 0.40

Fig. 3. The most similar fifteen tags expanded for given tag 'Eiffel' along with the similarity scores indicating the semantic similarity. We see that tag expansion using Google snippets effectively finds many related user tags that are difficultly to find with dictionary (or ontology) based methods.

2.1. Tag Expansion Using Google Snippets

Due to the sparsity of user-contribited tags in Flickr [7], we propose to expand associated tags of visually similar photos given the unlabeled image. The diverse and noisy tags for describing the same semantic topic makes the tag expansion vital and challenging as illustrated in Figure 2. We propose to expand user tags by using Google snippets, which are the text returned by search engine as brief information of websites when user conducts keyword search. To feature the semantic meaning of user tags, we first expand each tag to a fixed length vector by bag-of-words called *anchor terms*.

In order to obtain the anchor terms informative and meaningful to represent each tag, we take each tag in our dataset as a query to Google successively. The words extracted from returned snippets are filtered by stop-word list and stemmed to obtain final anchor terms. In our dataset, we finally use anchor terms of 91,004 words extracted from top 50 snippets returned by Google for each tag.

Once the anchor terms are collected, each tag in the dataset is expanded as a feature vector represented with the anchor terms. We send each tag as a query to Google to obtain top 50 snippets as described above. Then we count the frequencies of anchor terms in these snippets to construct the feature vector. The frequencies of anchor terms are normalized by each snippet first then are summed up and averaged to obtain the final feature vector. Since we have represented each tag in our dataset as a feature vector, we compute the *cosine* similarity between two tags to obtain the semantic similarity between them. We show a tag expansion example in Figure 3 for tag 'Eiffel.'

Finally, given an unlabeled image as input, we rely on established CBIR system (cf. Section 3.1) to retrieve visual similar photos for the unlabeled image. The expanded tags of these simiar photos are taken as candidate annotations T. We expand each tag with top N similar tags in our dataset. Empirically, we set N as 3 to get reasonable results and not to include too many noisy tags.

2.2. Candidate Annotation Ranking Using Visual and Semantic Consistency

Given a unlabeled image I_q, the candidate annotations T extracted above are mixed with relevant and irrelevant tags. In order to reject these irrelevant tags, we rank candidate annotations by estimating tag relevance $rel(I_q, w)$, linearly combining *visual consistency* $rel_v(I_q, w)$ and *semantic consistency* $rel_s(I_q, w)$. The intuition for visual consistency is that if a tag is relevant to the input image, the image retrieval results returned by text-based image retrieval using the tag should be similar to the CBIR search results with input image as the query. For semantic consistency, if a tag is relevant to input image, its similar tags should be relevant

to input image too. If not, the tag is probably noisy. We then linearly combine the relevance scores to rank the each candidate tag as $rel(I_q, w) = rel_v(I_q, w) + rel_s(I_q, w)$.

2.2.1. Visual consistency (rel_v)

Previous works [4, 6] have introduced some tag relevance estimation methods based on tfidf (term frequency inverse document frequency) weighting or frequencies solely for each similar photo from CBIR results given the unlabeled image. However, these methods are unreliable for Flickr photos generally with few and noisy user-contributed tags. Instead, we estimate tag relevance for whole tag candidates and further consider its relevance to the unlabeled image.

Intuitively, if we can exploit both textual and visual cues to jointly estimate the similarity between input image and a tag w, the tag relevance will be more robust. We introduce a novel method by comparing the retrieval results of CBIR and text-based[1] image retrieval. We observe that if the image list returned by CBIR for I_q is similar to the image list returned by text-based image retrieval using w as a query, then the tag w is more relevant to I_q. Given an unlabeled input I_q and tag w, the relevance score is computed as:

$$rel_v(I_q, w) = \sum_{J \in \Theta_q \cap \Theta_t} tfidf(w, J) \times sim_{vis}(J, I_q), \quad (2)$$

where Θ_q and Θ_t are the image lists returned by CBIR with I_q and text-based image retrieval with w, respectively. $tfidf(w, J) = freq_{w,K_J} \times \log \frac{N}{n_w}$, where K_J is associated tags of image J, $freq_{w,K_J}$ returns the frequency of w in K_J, N is the number of images in dataset, n_w is the number of images in which the tag w appears in associated tags. Apparently, rel_v considers both tag significancy (by *tfidf*) and visual similarity to the unlabeled image over the overlapped photos from both retrieved results.

2.2.2. Semantic consistency (rel_s)

We further improve the tag relevance estimation by including similar tags of tag w. Intuitively, if a tag has high relevance score with input image, its semantically similar tags (cf. Section 2.1) should be computed as highly relevant tags. If not, the tag is probably a noisy one and its similar tags obtained by tag expansion are not consistent in semantics. We measure the tag semantic consistency by summing up the visual relevance scores of top r similar tags of w as:

$$rel_s(I_q, w) = \sum_{x \in X(w)} rel_v(I_q, x), \quad (3)$$

where $X(w)$ is the top r similar tags of w (cf. Section 2.1). Empirically, the r in our experiments is set as 10.

To sum up, for annotating an unlabeled image, we first form the candidate annotations by choosing all expanded tags from its visually similar photos from CBIR search results. The tag relevance of each candidate annotation with respect to the input image is then computed using Eq. 2 and Eq. 3. All candidate annotations are ranked in descending order according to their relevance scores. Next we select top k ranked annotations as final annotations. Empirically, the common setting of k is 5. We also select top 5 ranked annotations in our experiments.

[1]We employ the MySQL full-text search model to retrieve photos with associated tags.

	Our method	SBIA_PRO	AnnoSearch
	**coffee beans** **coffee** **need my coffee** golden gate bridge tea	snow beach **coffee** winter cup	beach snow winter trip beach sand light
	**toureiffel** **la tour eiffel** **torre eiffel** **tour-eiffel** **eiffel-tower**	**paris** france snow **night** beach	**paris eiffel tower** **paris night** **tour eiffel** travel europe snow winter
	**arc de triomph** **arc du triomphe** **arc de triomphe** torre eiffel snowing	beach snow **paris** france travel	snow winter beach ocean **france paris** eiffel tower paris sea

Fig. 4. The annotation results for three test images. Prior methods (e.g., SBIA_PRO, AnnoSearch) mainly consider frequency counts for the tags in CBIR return results and are dominated by frequent tags such as "france paris," "beach," etc., in the last image; however, our method can select those salient tags ranked by visual and semantic consistency.

3. EXPERIMENTS

3.1. Experimental Setup

Dataset: As the dataset for automatic image annotation, we sample a subset of Flickr550 dataset [8]. The subset contains 3,378 images manually labeled across seven categories: beach (1,657), beer (365), coffee (343), eiffel tower (554), golden gate bridge (192), pisa tower (144), triomphe (123). In addition to these labeled images, we add other noisy Flickr images into the subset to construct a 8k Flickr image subset. Besides, we have sample 10 images per category for the seven categories as our test set.

Visual features: We have established a multimodal CBIR system for image retrieval. To balance global features and local appearance, as the authors suggest [8], the visual modalities include: 225-dimensional grid color moment, 48-dimensional Gabor texture, and 3500-dimensional visual words. The retrieval results of three modalities are then normalized and fused in an average manner. We adopt KD-tree to index these visual features for on-lne retrieval.

Evaluation metrics: Due to no available ground truth for automatic image annotation under unlimited vocabularies, we measure the performance of automatic annotation by three metrics as [7]: *mean reciprocal rank* (MRR), *success at rank 1* (S@1) and *precision at rank 5* (P@5). For the annotations generated by various methods, human reviewers are asked to mark them as relevant or irrelevant. The judged results are then evaluated for their MRR, S@1 and P@5.

3.2. Automatic Image Annotation Results

In order to evaluate our proposed method, we have compared it with three previous search-based image annotation (SBIA) methods: AnnoSearch [3], SBIA with prominence (SBIA_PRO) and if-ikf (SBIA_IFIKF) scores [4]. In summary, prior SBIA methods to compute a tf-idf-like term frequency score for a tag that is normalized by the size of tags along within each retrieved photo. They are unreliable due to very sparse photo annotations and also confirmed in the experiments.

In these experiments, we retrieve top 100 visually similar images for an input image by CBIR and generate top 5 ranked annotations by each method. The overall performance is shown in Table 1. Because of the unavailability of initial keywords associated with input

Table 1. Annotation performance of different methods and performance metrics averaged over 70 test images (in 7 categories).

	MRR	S@1	P@5
AnnoSearch [3]	0.43904	0.28571	0.29714
SBIA_PRO [4]	0.44786	0.32857	0.20857
SBIA_IFIKF [4]	0.42619	0.31429	0.23429
Our method	**0.52429**	**0.38571**	**0.37429**

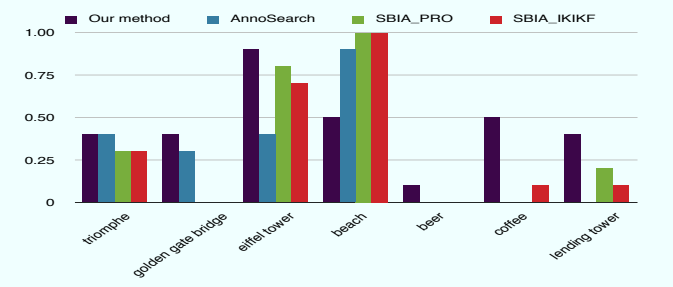

Fig. 5. The performance comparison for different methods in S@1 metrics across categories.

images, we ignore the initial keywords used in AnnoSearch implementation [4].

Table 1 shows that our method outperforms the prior ones, which mainly utilize frequency counts of the surrounding tokens (tags) for the CBIR-retrieved Flickr photos. Our method relatively improves 19%, 23% and 17%, respectively for AnnoSearch, SBIA_IFIKF and SBIA_PRO in MRR. Since MRR is sensitive to the rank of the first relevant tag in rank list, the improvement indicates that our method tends to give relevant annotations higher ranks. For S@1, our method has relative improvements as 35%, 17%, and 23%, respectively for AnnoSearch, SBIA_PRO, and SBIA_IFIKF. The improvements show that if we only consider the top 1 ranked annotation for input image, our method generates a relevant annotation in higher probability than other methods. P@5 measures the proportion of relevant annotations in the rank list. Our method still performs better than the prior three ones with the relative improvements of 26%, 79%, and 60%. It indicates that our proposed method gives more relevant annotations by tag expansion and considering visual and semantic consistency.

For further investigating the performance breakdown, we have showed Figure 5 and Figure 6 for S@1 and P@5 results. In both Figure 5 and Figure 6, our method competes or outperforms other methods in all categories except for "beach". We find there are 1,657 manually labeled "beach" in our 8k Flickr image subsets. The large proportion of "beach"-annotated images favors the three previous methods because these methods heavily rely on tag frequencies to estimate tag relevance and mostly ignore visual similarities between tag and images pairs and semantic consistency between recommended tags. What interests us in both figures is that for the object-based categories such as "coffee" and "beer," our method has a reasonable annotation performance compared to three prior methods that almost fail on these categories. We show several automatic image annotation results in Figure 4 for demonstration.

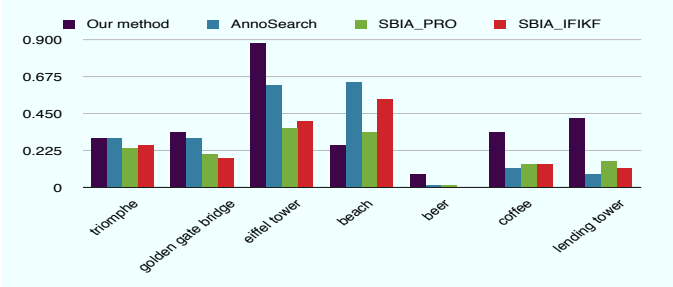

Fig. 6. The performance comparison for different methods in P@5 metrics across categories.

4. CONCLUSIONS AND FUTURE WORKS

Search-based approaches promise an automatic image annotation solution with unlimited vocabularies and preventing tedious training processes. We argue to leverage rich context in growing user-contributed photo collections for search-based automatic image annotation. To remedy the sparsity and noise problem in user-contributed tags, we propose to expand user tags by Google snippets and further improve tag ranking by novelly considering both visual consistency and semantic consistency among candidate tags. Evaluations over large consumer photos show that our proposed method outperforms prior works in many aspects. In future work, we plan to extend the experiments for larger consumer image datasets and compare the utilities among auxiliary knowledges from Web images and Flickr photos.

5. REFERENCES

[1] Liangliang Cao and et al., "Annotating photo collections by label propagation according to multiple similarity cues," in *ACM Multimedia*, 2008.

[2] Li-Jia Li and Li Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *CVPR*, 2007.

[3] Xin-Jing Wang and et al., "Annosearch: Image auto-annotation by search," in *CVPR*, 2006.

[4] Changhu Wang and et al., "Scalable search-based image annotation of personal images," in *MIR*, 2006.

[5] Xirong Li, Cees G.M. Snoek, and Marcel Worring, "Learning tag relevance by neighbor voting for social image retrieval," in *MIR*, 2008.

[6] Xirong Li and et al., "Annotating images by harnessing worldwide user-tagged photos," in *ICASSP*, April 2009.

[7] Börkur Sigurbjörnsson and et al., "Flickr tag recommendation based on collective knowledge," in *WWW*, 2008.

[8] Yi-Hsuan Yang and et al., "Contextseer: Context search and recommendation at query time for shared consumer photos," in *ACM Multimedia*, 2008.

[9] Lei Wu, Linjun Yang, Nenghai Yu, and Xian-Sheng Hua, "Learning to tag," in *WWW*, 2009.