SheepDog – Group and Tag Recommendation for Flickr Photos by Automatic Search-based Learning

Hong-Ming Chen, Ming-Hsiu Chang, Ping-Chieh Chang,
Ming-Chun Tien, Winston H. Hsu, and Ja-Ling Wu

Communications and Multimedia Lab
National Taiwan University
{blacksmith,cmhsiu,pingchieh,trimy,winston,wjl}@cmlab.csie.ntu.edu.tw

ABSTRACT

Online photo albums have been prevalent in recent years and have resulted in more and more applications developed to provide convenient functionalities for photo sharing. In this paper, we propose a system named *SheepDog* to automatically add photos into appropriate groups and recommend suitable tags for users on Flickr. We adopt concept detection to predict relevant concepts of a photo and probe into the issue about training data collection for concept classification. From the perspective of gathering training data by web searching, we introduce two mechanisms and investigate their performances of concept detection. Based on some existing information from Flickr, a ranking-based method is applied not only to obtain reliable training data, but also to provide reasonable group/tag recommendations for input photos. We evaluate this system with a rich set of photos and the results demonstrate the effectiveness of our work.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process;
H.3.5 [**Online Information Services**]: Web-based services

General Terms

Algorithms, Design

Keywords

Flickr, Concept Detection, Recommendation System.

1. INTRODUCTION

The number of digital images has exploded with the proliferation of digital photo-capture devices. At the same time, the progress of Internet and free online storage space have merged more and more online public photo sharing websites developed for photo amateur, such as Flickr[1]. People upload their photos to the websites, and thus a great many public consumer photographs are available online. Most users prefer their photos to gain public attention for social purposes [2], hence they usually add their photos to suitable photo groups one by one, and attach tags to each photo manually. Thus the browsers can easily find these photos by searching groups or tags. However, repeatedly adding photos to certain groups and attaching tags to each photo is quite exhausting. And it would be troublesome for general users since they do not know how many related groups exist, and which group they should add to. For example, there are about 15000 groups related to "dog" with different popularities on Flickr. If users deal with group selection manually, they will never make the best choice. Indeed, they need an easy-to-use tool to manage their photos in online albums.

Leveraging the power of Flickr service, many people use the Flickr API to design helpful tools. For example, given a keyword, both the *Findr* and the *Tag Browser* can find the most relevant tags and most relevant photos on Flickr. However, these tools still cannot solve the problems mentioned above. Wang *et al.* [3] proposed the *AnnoSearch system*, which provides a sophisticated mechanism for mining related tags of a given photo. Although this work can solve the tag attachment problem, it doesn't handle the issue about suitable group recommendation. And even if the system could annotate photos with appropriate tags, users still need to attach at least one accurate keyword for each photo to do *AnnoSearch*. If users want to cope with a bunch of photos at a time, they still have to struggle with keywords labeling.

Based on above observation, we designed the [1]*SheepDog* system to solve this problem. The user interface of our system is shown in Fig. 2. Users can once upload one or many photos to the *SheepDog*. Our system will return related popular groups/tags (Fig. 2 A, B) to the users in a few seconds. The system either can add these groups/tags to the photos automatically without manually intervener or let the users choose the groups/tags which they appreciate from our recommended lists. The framework of *SheepDog* is showed in Fig. 1. Fig. 1(a) shows the concepts that we defined for concept detection (as other standard machine learning works do). Fig. 1(b) shows the training data collection from Flickr. We obtain our training data by two comparative methods: *photo-level* mechanism and *group-level* mechanism. Then we extract their visual features, and use support vector machine (SVM)[4] to learn a probability-based model for all concepts(Fig. 1(c)). (In Lin's tool we use *probability_estimates* mode: train a C-SVC model for probability estimates) We can use this trained model to predict potential concepts for each test photo (Fig. 1(d)). Finally, we choose the top-n concepts from the prediction results, and use these results as keywords to mine the related groups and tags in Flickr. After applying our ranking algorithm, the system recommends suitable and popular groups/tags to the users (Fig. 1(e), (f)).

MM'08, October 26-31, 2008, Vancouver, British Columbia, Canada.

[1] *Sheepdog is a type of domestic dog whose original mission was to herd sheep to their own group. The term "SheepDog" is a metaphor here, because we want to herd our photos to their suitable groups.*

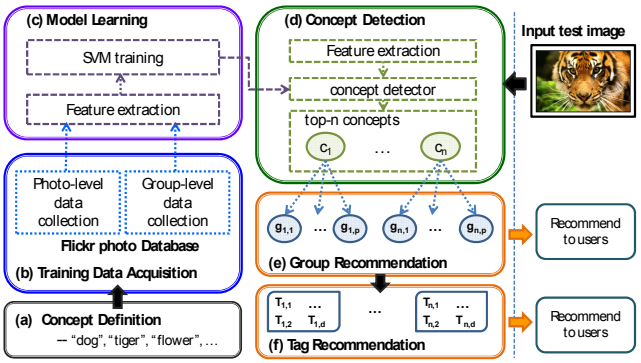

Fig. 1 The overall framework of *SheepDog*. In this system, the concept model is constructed based on web searched training data. The primary contribution of this work is a system for choosing search-based model either by photo-level or group-level data acquisition; Given an input photo, the system predicts possible concepts and automatically recommends reliable groups and tags for this photo.

Kennedy *et al.*[5] explored the trade-offs in acquiring training data for concept detection through automated web search as opposed to human annotation. However, they only compared human annotations result with their search results. Another significant contribution of our work is that we provide a brand new idea of "how to acquire reliable search-based data" for machine learning. We compare two source of *pseudo-positive* photos acquisition in our experiment to demonstrate our idea, including "photo-level-search" results and "group-level-search" results on Flickr. From the experiments, we find that group-level-search has much better performance than the other, which provides a remarkable result for future research.

This paper is organized as follows. Section 2 discusses the data acquisition and prediction system. Section 3 presents the recommendation mechanism. Section 4 gives experimental results of our system, and Section 5 concludes this work.

2. CONCEPT DETECTION

There are numerous possible concepts provided on Flickr, in order to properly demonstrate our idea, we define the scope of adopted concepts. We empirically define 62 general concepts which appear frequently in consumer photographs. These 62 concepts can be put into several categories including "animal", "architecture", "nature scene", "portrait", "plant", and "color-oriented" roughly. The complete concept names and experimental results are shown in: http://u.csie.org/sheepdog .

From the perspective of web-search methods, we collect reliable training data for each concept with two mechanisms: *photo-level* and *group-level* which are described in Section 2.1 and 2.2 respectively. Several visual features are extracted to describe each training photo, including color, texture, and edge features[6]. We choose these three kinds of features because they have been shown to work well for image classification problem in the past. After fusing these extracted features, a 369 dimensional feature vector formed for each photo. With the aid of SVMs, the mapping between low-level features and high-level concepts could be learned. Note that our case is not a "yes or no" problem. What we really want to know is the degree that a test photo fits each of our defined concepts, rather than whether a test photo belongs to a specific concept or not. For a given photo, it may fit multiple

Fig. 2 SheepDog - The web interface of the groups/tags recommendation system. A and B are our system recommendation lists (correspond to Fig. 1 (e), (f) respectively) for the input photo C that user selected. Besides our recommendation result, users also can choose the tags/groups which they appreciate from them.

concepts, e.g., a photo with a "dog" playing with a "cat" on the "grass". If these concepts have been defined in our concept set, the SVM predictor should give these appeared concepts higher probabilities. Once we have the SVM prediction results, we can select concepts having higher probabilities to do group/tag recommendation. The recommendation procedure will be described in Section 3.

To discover "how to acquire reliable training data by search-based methods," we design two sophisticated mechanisms and compare the results in Section 4. The first mechanism is photo-level search method, and the other one is group-level search method.

2.1 Training data acquisition by photo-level search

In *photo-level* search, to collect training photos of a specific concept, we directly use the concept name as keyword to do *tag search* on Flickr. The search result shows all photos tagged with the keyword. However, as mentioned in [5], the *tag search* doesn't provide very accurate result. Going a step further, the Flickr can sort the *tag search* result by the *interestingness* of each returned photo, which is evaluated by Flickr's proprietary method. Flickr uses quantities related to user behaviors such as the *viewed times, comments, set as favorite* to evaluate a photo's *interestingness*. We believe that *interestingness* is reliable information because to some extent, the most interesting photos in the search result are more representative and with less misleading. Furthermore, we observe that the semantic meanings of these most interesting photos are very close to the search keyword, i.e. the tag or the concept name. We take the top-*K* (i.e. 250) most interesting photos for each concept as our training data. The training data obtained from our *photo-level* mechanism have much less noise than only using tag search result.

2.2 Training data acquisition by group-level search

In the *group-level* search, we also use the concept name as keyword to collect the training data of a specific concept. However, we do *group search* instead of *tag search* on Flickr. The search result shows the related groups sorted by *relevance*, which is evaluated by Flickr itself, too. The returned amount of related groups is huge, taking the concept "dog" as an example, about 15000 groups are returned. Some of these groups are popular groups with lots of photos inside and having many group members, while some are rarely visited. In order to obtain reliable training data, we have to select representative groups from the *group search* result. The groups with longer history, more members and more photos, would be more reliable. Hence, we consider these four factors which are obviously important to decide whether a group is representative of a specific concept. We use the Borda rank [7] to fuse the four factors and sort these groups. The higher ranked groups are more representative. Then we select the top-*L* (i.e. 10) groups from the ranking result. In each of the selected group, we sort all photos inside the group by *interestingness*, just as what we do in *photo-level* search. The amounts of the most interesting photos in each group are picked proportional to the ranking result – the higher rank a group is, the more photos we will select from it.

3. RECOMMENDATION

With the technique described in Section 2, we collect reliable training data and construct the statistical model for each concept. In this section, we introduce the recommendation algorithm to recommend appropriate groups and tags for an input photo.

3.1 Group Recommendation

To accurately recommend appropriate groups, the system first utilizes the trained model to predict the top-*n* concepts of an input photo *I*. There are 62 concepts in our experiments, in the prediction step, the SVM predictor gives each concept a probability value to indicate the degree that the photo *I* fits this concept. We formulate the probability distribution as $p_I = (p_1, p_2, \ldots, p_{62})$. The sum of all probability values over these 62 concepts is 1 for each input photo. Probabilities of each concept are then ranked in descending order. The top-*n* concepts are extracted to be the best matched concepts of *I*. This procedure is illustrated in Fig. 1(d).

Similar to the representative group selection step described in Section 2.2, the name of each of the top-*n* concepts is taken as a keyword to do *group search* on Flickr. For each *group search* result, we consider the four identical factors to obtain the top-*p* (i.e. 5) highest scored groups, which are more representative and popular than other groups. Consequently, the system recommends $p \times n$ suitable groups for photo *I*, as shown in Fig. 1(e).

3.2 Tag Recommendation

In the tag recommendation stage, the system automatically attaches proper tags for an input photo *I*. Since representative and popular groups usually accompany with high quality and popular tags at the same time, we choose tags from these groups for recommendation. For each of the top-*n* concepts predicted for a photo *I*, we recommend top-*p* suitable groups, and discover popular tags by gathering the statistics of the tags which was ever attached to any photo in these groups. The system selects tags which are used more than *d* (i.e. 2) times as tag candidates. This procedure corresponds to Fig. 1(f). We provide a user friendly interface for users to select their favorite groups and tags from the recommended list. We have demonstrated the interface in Fig. 2 and give a using scenario in introduction.

4. EXPERIMENTAL RESULTS

This section presents the experimental results to evaluate our framework. Since our recommendation mechanism mainly depends on the result of concept detection, we evaluate the accuracy of concept detection by two experiments in this section. We also compare the results of the proposed two training data collection mechanisms.

4.1 Concept Detection Evaluation

To evaluate the accuracy of concept detection, we collect 50 photos for each concept from Flickr. All these photos are examined by human to ensure they do match this concept, and we regard these photos as ground truth. Since we define 62 concepts in our experiments, there are totally 3100 photos in the test data set. First, we evaluate the detection result by considering if a photo is first classified into its own human annotated concept. For example, given a photo annotated with "tiger", we examine whether the "tiger" concept have highest predicted probability. Over the 3100 test photos, our SVM prediction result shows that the average hit rates are about 18% and 22% for the *photo-level* and the *group-level* mechanisms, respectively. Second, we also calculate the average probability distribution $\overline{p}_c$ over the 50 photos for each concept to show the trend of the detection result, where $\overline{p}_c$ is defined as:

$$\overline{p}_c = \frac{1}{N}\sum_{i=1}^{N} p_i \text{ , (N = 50),} \qquad (1)$$

where $p_i = (p_1, p_2, \ldots, p_{62})$. Hence there are 62 average probability distributions, each shows the prediction trend for one concept. In each distribution, we sort the probability values in descending order. Take the concept "tiger" as example, the sorted average probability distribution is illustrated in Fig.3. Both the *photo-level* and the *group-level* training data collection method result in very well prediction trend because the "tiger" itself has highest probability, and the other top-5 are all concepts in the "animal" category.

Give an observation to these 62 average probability distributions, we find that almost all distributions have the trend to first classify its own concept *C* to top-1, while the other recommended top-5 concepts are almost in the same category which contains concept *C*. Furthermore, the *group-level* method has better result than that of the *photo-level* method. Given a concept *C*, we can obtain high related concepts according to the experiment results. However, a photo containing concept *C* will not always contain these high related concepts. Moreover, this experiment result cannot reflect the diversity of each test photo. To evaluate how well the *SheepDog* can predict concepts for each photo, we invited several users to do the subjective test.

4.2 Subjective Test

In the first experiment, all the ground truth photos are annotated with only one concept. However, one photo usually contains more than one concept. Hence, we designed another experiment to evaluate the performance of concept detection based on subjective test. We randomly choose 930 photos (15 photos for each of the 62 concepts) from these 3100 test photos. These recommendation

Table 1. Average score of photos in each category of the 62 concepts. Here we can give a quick conclusion that group-level based search works better than photo-level based search, and in the top-3 system recommended concepts, both have more than 1.5 concepts that users appreciate in Overall Average.

	Animal	Architecture	Nature Scene	Portrait	Plant	Color Oriented	Overall Average
Photo level	1.10	1.51	1.72	1.07	1.79	1.51	1.55
Group level	1.24	1.68	1.90	1.40	1.92	1.56	1.69

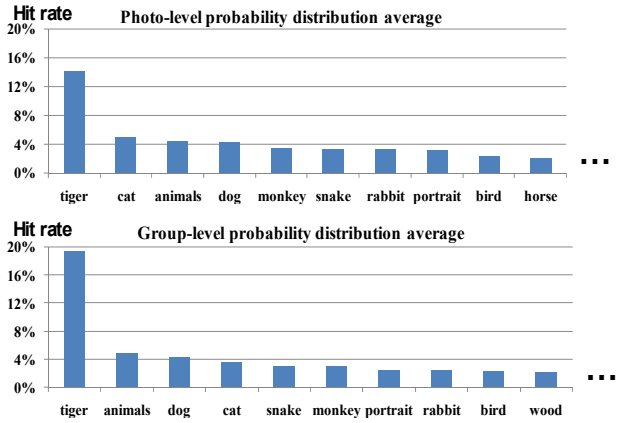

Fig. 3 Average probability distribution with top-10 prediction result of the "tiger" concept. The top and the bottom charts show the detection result of photo-level method and the group-level method, respectively. Both charts show the top-10 concepts with highest average prediction probabilities. Both two have good result to classify photos into correct or similar concepts. And the group-level method has better result to classify photos into the correct concept.

results are scored by 15 persons. We evaluate the scores based on [5]: annotate each recommended concept with "Perfect", "Correct", or "Wrong". Consequently, for each concept, we obtain an average score S to indicate how the system predicts this concept:

$$S = (P*1 + R*0.5 + W*0)/N_c \quad (N_c = 15) \qquad (2)$$

Where N_c denotes the photo numbers for each concept. P, R, and W are the number of "Perfect", "Correct", and "Wrong" annotations respectively. Empirically, most consumer photographs contain 2~3 concepts in each photo, so we demand the users to evaluate the system's top-3 recommended results. Therefore, by Eq. 2, the score of a concept will ranges from 0 to 3.

The Eq. 2 is improved from [4], so that the score can directly indicate that how many concepts that users appreciate in our top-3 recommendation. In Table.1, for each category, we average scores of all concepts inside. The results vary because of the different property of each category. Categories with obvious features are prone to have higher score. We also obtain an overall average result (in the last column) by averaging all 62 concept scores to show the overall performance. Both the *photo-level* and the *group-level* methods have overall average scores higher than 1.5, which seems good results that the user can acquire more than one perfectly matches among the 3 recommended concepts.

In this subjective test experiment, we also compare the recommendation results of *photo-level* method and *group-level* method in Table.1, and the group-level method still works better. We figure out an explanation to this observation. A group associated with concept C would actually correlate to this concept - people who put their photos into this group implies that they annotate the concept C as ground truth for their photos. Therefore, photos in this group would also closely correspond to concept C. However, for a photo tagged with C, the concept C may only be a minor concept of this photo. In other words, the concept C might not really match the main concept of the photo. [5] also mentioned that when photos are tagged by Flickr users, there is only about 50% chance that the concept actually appears in the photo.

5. CONCLUSIONS

In this paper, we proposed a reliable system for automatically adding photos into proper and popular groups. In addition, the system recommends suitable tags for photos and provides a user friendly interface such that users could easily select their favorite tags to attach. We also design two methods to collect training data for concept detection based on the idea of web search. Both the *photo-level* search and the *group-level* search contain the ranking mechanism to obtain representative training photos for each concept. The experiments compare the results of the two training data gathering methods and show the effectiveness of our concept classification and group/tag recommendation approaches.

6. REFERENCES

[1] Flickr, http://www.flickr.com/

[2] M. Ames and M. Naaman, "Why We Tag: Motivations for Annotation in Mobile and Online Media," in Proc. SIGCHI CHI'07, 2007.

[3] X. Wang, L. Zhang, F. Jing, and W. Y. Ma, "AnnoSearch: Image Auto-Annotation by Search," in Proc. IEEE CVPR'06, vol. 2, pp. 1483-1490, 2006.

[4] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[5] L. S. Kennedy, S.-F. Chang, and I. V. Kozintsev, "To search or to label?: predicting the performance of search-based automatic image classifiers," in Proc. ACM MIR'06, pp.249 – 258, 2006.

[6] A. Yanagawa, W. Hsu, and S.-F. Chang, "Brief Descriptions of Visual Features for Baseline TRECVID Concept Detectors," Columbia University ADVENT Technical Report #219-2006-5, 2006.

[7] Mc Donald K and A.F. Smeaton, "A Comparison of Score, Rank and Probability-based Fusion Methods for Video Shot Retrieval," CIVR 2005.