Tag Suggestr: Automatic Photo Tag Expansion Using Visual Information for Photo Sharing Websites

Onur Kucuktunc, Sare G. Sevil, A. Burak Tosun, Hilal Zitouni, Pinar Duygulu, and Fazli Can

Bilkent University, Department of Computer Engineering, Ankara 06800, Turkey

Abstract. In this paper, we propose an automatic photo tag expansion system for the community photo collections, such as Flickr[1]. Our aim is to suggest relevant tags for a target photograph uploaded to the system by a user, by incorporating the visual and textual cues from other related photographs. As the first step, the system requires the user to add only a few initial tags for each uploaded photo. These initial tags are used to retrieve related photos including the same tags in their tag lists. Then the set of candidate tags collected from a large pool of photos is weighted according to the similarity of the target photo to the retrieved photo including the tag. Finally, the tags in the highest rankings are used to automatically expand the tags of the target photo. The experimental results on Flickr photos show that, the use of visual similarity of semantically relevant photos to recommend tags improves the quality of suggested tags compared to only text-based systems.

1 Introduction

Recently, large number of photos have become available in photo sharing services. Although the advances in content based image retrieval studies are promising [12], scaling these techniques to web is difficult. On the other hand, users are willing to annotate the images manually [8] allowing the tag based search systems to be practical.

In the community photo collections, such as Flickr, tags are generally assigned by users who upload the photos, identifying the location (place, country, etc.) where the photo is taken, as well as the objects/people appearing in the image, together with some specific words related to camera characteristics, interest groups etc. However, the tags are usually subjective, noisy and in a limited number, reducing the accessibility of the photographs. Tag suggestion systems, that can provide related tags to be selected, are therefore important to eliminate the limitations and to guide the users.

Our motivation in this work is to provide a system that enhances Flickr's search capabilities by automatically recommending meaningful tags for annotating photographs. We propose a tag suggestion system which expands the tags

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

D. Duke et al. (Eds.): SAMT 2008, LNCS 5392, pp. 61–73, 2008.

of an image provided by the user, incorporating the tags of the other images which are visually similar. The main contribution of our approach lies in the use of visual information, unlike the previous studies which focuses only on textual information.

First, a few already existing tags (in the order of 2 or 3) are used to form an initial query to find related images including these tags. Then, all the other tags co-occurring with these images are listed as the candidate recommendations. The tags are then weighted according to the similarity of the images, resulting in a higher ranking for the tags coming from visually similar images.

The rest of the paper is organized as follows: related work is discussed in Section 2, proposed tag suggestion algorithm is explained in Section 3, experimental work and evaluation techniques are presented in Section 4, results are discussed in Section 5, and finally, Section 6 includes conclusions and possible future work.

2 Related Work

Automatic and semi-automatic annotation of photos has been widely studied throughout the years. Many studies in the field use text based, probabilistic and frequency oriented methods which all have their own restrictions.

Elliot and Ozsoyoglu describe a method for semi-automated semantic digital photo annotation in [9]. Related concepts, keywords, time and location information of a target photo are used for generating a set of related photos, and their tags are ranked according to a scoring function. Naaman *et al.* [1] use a context based approach for the annotation of persons in a photo. Their method requires time and space information when each photo is taken. Although, with wide usage of digital cameras, time information can easily be retrieved, space (i.e. physical location) information can only be obtained from a system with GPS support. Yan *et al.* [2] categorize tags in two categories: some of them are used in browsing and the others are used in tagging. The tags that are chosen for browsing purposes are not suggested to the users; only the remaining set of tags can actually be used for annotation purposes. To make this categorization, the method makes a frequency-based analysis and words that are widely used are chosen as good candidates for making efficient and effective browsing operations. The less frequently encountered words however, are seen to have discriminative properties so they are used in actual tagging.

There are not many work on annotation suggestion methods that use both keyword-based and visual feature based similarities. Wenyin *et al.* propose a progressive semi-automatic image annotation strategy that use keyword-based and content-based image retrieval and relevance feedbacks in [3]. They claim that manual annotation is a tedious but very accurate process, since tags are selected based on human determination of the semantic content of images. This strategy is used in *MiAlbum* system [4] and evaluations show that it is effective for annotating images in photo databases. The system annotates multiple photos when a query is given with the user's feedback. Similarly, Suh and Bederson describe their approach for efficient bulk annotations in [5,6], and they create meaningful

image clusters for this purpose. However, an automatic approach is desired in our case, and we need to suggest tags for one photo. Therefore, annotation of groups of photos by the given strategies does not solve our annotation problem.

It is generally thought that computer vision techniques create a heavy overload and textual methods can be enough to provide good systems and thus they are not prefered to be used by researchers. But, though they are still developing, vision based methods are quite powerful and when combined with textual methods, very effective automated systems can be achieved. A good application of combined use of textual and visual techniques is proposed by Quack *et al.* in [11]. Objective of the work persented in [11] is to provide a system that automatically forms high quality image databases using the large-scale internet sources. They retrieve large numbers of raw data consisting of geotagged images togehter with their corresponding associated information(which include tags, title, description, time stampts etc.). They use textual, visual and spatial information to cluster these images. Then they classify their clusters into 'objects', which they define to be physical items on fixed locations, and 'events', special social occasions taken place at certain times. Using these specific classes in formed clusters, they associate images to wikipedia articles and check the validity of these associations. The final output of the system then becomes nicely organized groups of images with relevant enclopedia information attached to them.

We have observed that pure text-based approaches cannot provide perfect systems, as the visual content is totally independent from the textual content. Therefore the proposed method uses the advantages of both visual and textual techniques for obtaining high performance.

3 Tag Suggestion Algorithm

Our system is a stand-alone application that serves as an interface for uploading photos to Flickr. Main purpose of the system is to recommend tags to users as a photo is being uploaded, so that the probability of entering irrelevant tags to Flickr is reduced. To do this, user is required to provide initial tags with which the system retrieves related photos. Recommended tags are chosen among the distinct tags that come along with the set of related photos. While recommending a tag, visual similarities between a related photo and the photo to be uploaded are taken into account.

Figure 2 visually describes the proposed method. Algorithm steps are explained in further detail in the following subsections.

The method can be summarized in the following steps:

1. Obtain target photo and corresponding initial tags from user. Let I_t be the target photo to be uploaded, and $T_{init} = \{t_{init1}, t_{init2}\}$ be the initial tags for this photo.
2. Connect to Flickr server and fetch the first m relevant photos $I_R = \{I_1, ..., I_m\}$ (and their corresponding tags) $T(I_i)$ containing the given initial tags.

$$\forall I_i \in I_R,\ T_{init} \subset T(I_i) \tag{1}$$

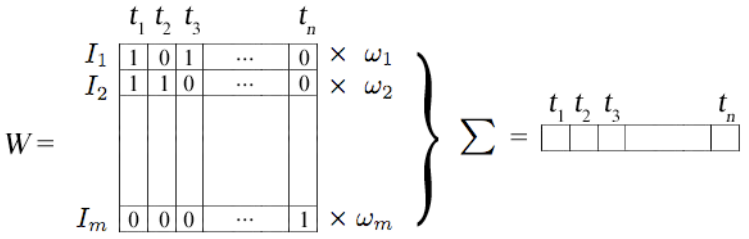

Fig. 1. Calculation of total weights W for each distinct tag in T

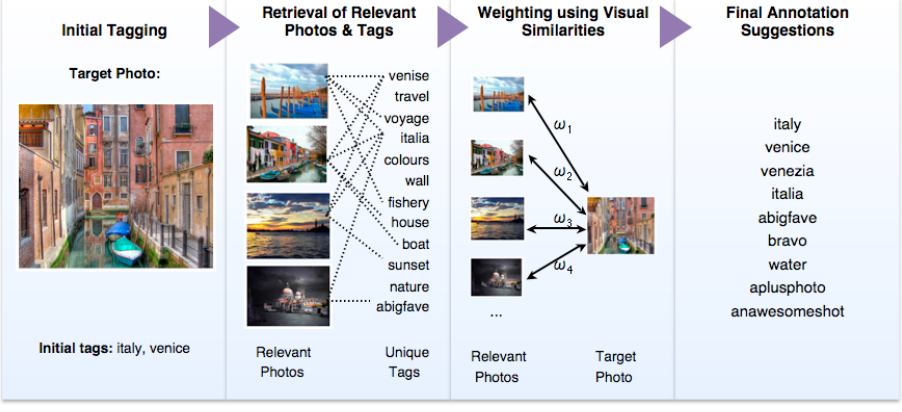

Fig. 2. Overview of the proposed method

3. Let $T = \{t_1, t_2, ..., t_n\}$ be the unique set of tags of all relevant photos, which contains n distinct tags.

$$\forall I_i \in I_R,\ T(I_i) \subset T \tag{2}$$

4. Extract visual feature f_{I_t} for the target photo, and f_{I_i} for all relevant photos.
5. Find the weight ω_i representing the visual similarity between target photo and the i$^{\text{th}}$ relevant photo I_i as Eqn 3.

$$\omega_i = \frac{1}{dist(f_{I_t}, f_{I_i})},\ i \in \{1, ..., m\} \tag{3}$$

6. Generate a binary $m \times n$ matrix C, (where n is number of unique set of tags, m is the number of relevant photos). Set (i, j), if photo I_i contains tag t_j.

$$C_{ij} = 1 \Leftrightarrow t_j \in T(I_i) \tag{4}$$

7. Multiply each row i with the visual similarity ω_i, sum the columns to get a $1 \times n$ matrix W of tag weights as shown in Fig 1 where

$$W_i = \sum_{j=1}^{m} C_{ji} \times \omega_j \tag{5}$$

8. Suggest tags in T according to their total weights W in non-decreasing order.

3.1 Initial Tagging and Retrieval of Related Photos from Flickr

In order for the system to be able to recommend tags, users need to provide a photo to be uploaded, together with a number of *initial tags*. These initial tags are the ones that need to be given by the user without any restrictions coming from the system. Although initial tags can be chosen freely, choosing the most descriptive tags as initial ones is encouraged for effective results.

The purpose of the set of initial tags is to retrieve a set of photos from Flickr that have a higher probability of being related to the target photo to be uploaded. When retrieved, recommended tags are chosen among the tags of this set of related photos. We define a *related photo* to be a photo that contains all of the initial tags provided by the user. In other words, the initial tags are AND'ed in order to obtain these photos from Flickr. As mentioned before, once identified, both the complete list of tags and the photos themselves are fetched from Flickr.

The main assumption of our recommendation system is that the tags initially given by a user has a high probability of being used in photos *similar* to that photo. Therefore, after retrieving the set of related photos and their corresponding tags, a new set containing distinct tags is formed. In the following steps, weights are given to these tags with respect to the similarity factor between the target photo and photo(s) that contain that tag. Finally, tags with higher weights are recommended to the user.

3.2 Visual Feature Extraction and Similarity Calculation

Finding visual similarities between our target photo and related photos that are retrieved from Flickr is a very crucial step for our method, therefore visual features are needed to be extracted from each image. There are a number of alternative approaches for selecting and implementing visual features. In our method two common and effective visual features, namely color histograms and interest points, have been implemented for evaluation.

For the color histograms, we considered two color spaces; the RGB and the HSV color spaces and found the results from HSV color space to be more effective. Using 8 bins for each band, we obtained a feature vector of length 24 for each image and we calculated the similarity between a pair of images by calculating the Euclidean distance between the feature vectors.

For the finding the interest points, the SIFT operator, [10], is used. From these interest points, similarities between image pairs are calculated by using the matching algorithm provided by Lowe [10]. The total number of match points

between compared images are used as the similarity measure, and this value is normalized by the number of keypoints in target image.

3.3 Final Tag Suggestions

As mentioned before, a scoring function is applied to our list of candidate tags. We have used the visual features and their corresponding similarity measures separately to evaluate both approaches. However, in both cases, our scoring functions compute a weighted sum of the similarity values. After all candidate tags have been assigned a weight, tags with highest total weights are suggested to the user.

4 Experimental Work

To experiment on our work, we gathered 100 randomly chosen target photos from Flickr. Among these 100 photos there were some that were not suitable for our tests due to having too few number of tags or too specific tags (the necessity of these parameters will be clarified in section 4.3). Unsuitable photos were eliminated and performance measurements were made on the outputs of the remaining 66 target photos. In order to make a complete evaluation of the system, all tags of these set of target photos were analyzed and initial tags were chosen manually. As a design choice, size of the related photo set was selected as 100, thus about 7000 photos have been processed throughout the experiments. The following subsections describe experimenting environment, experiment results and evaluation methods are in further detail.

4.1 Experimental Environment: Flickr

As it has been mentioned before, proposed system is specifically designed to be used for the web site Flickr. With hundreds of millions of photos and over eight million users, Flickr is a rapidly developing web site that has a high potential of becoming a good source to be used by researchers working on social networking and content based image retrieval. Perhaps the most important reason for this rapid growth of attention is Flickr's emphasis on tagging. Through the use of tags, Flickr provides an image-browsing environment with various capabilities. As it is stated in Marlow *et al.*'s work in [7], Flickr has the following characteristic properties:

- *user-contributed* resources where users provide photos,
- *self-tagging* restrictions in which users can only tag the photos they have uploaded,
- *blind-tagging* behavior in terms of tagging support; tagging user cannot view other tags and the system does not suggest any possible tags to the user.

Tagging characteristics of Flickr are further discussed in [8]. According to the studies, although some photos contain more than 50 tags, statistics show that

photos with 1 to 3 tags covers more than 60% of all photos. Most frequent category types of these tags are locations, objects, people, actions, and time.

We have implemented our system using Java programming language with the Flickr API[2], which is available for non-commercial use. Flickrj[3], a wrapper library for Flickr API, was used for querying the database. According to Flickr APIs Terms of Use agreement, after color features and invariant points are extracted from a photo, we do not cache Flickr's data.

4.2 Choosing Optimal Number of Initial Tags

For our experiments, we first examined the optimal number of initial tags to be specified. In our proposed method, users provide initial tags for the photo to be uploaded, and then we retrieve the contextually relevant photos in Flickr by using these initial tags. The system needs to retrieve about 100 relevant photos per target photo, so initial tags should be selected carefully. First of all initial tags should not be too specific as they would not return sufficient number of related photos. Second important factor is the number of initial tags to be used. When few number of general tags are chosen, we end up having thousands of relevant photos, of which only a small portion is used. On the other hand, as the number of initial tags increases, we get fewer and more specific photos. In this case, the number of relevant photos may not be enough for effective recommendations. From our studies we have found that using 2-3 initial tags gives the best results considering for the proposed method. The table of average number of relevant photos in Flickr for a given number of initial tags are given in Table 1.

Table 1. Average number of relevant photos for a given number of initial tags

Number of initial tags	Number of relevant photos
1	514044
2	4050
3	95
4	14
5	4

4.3 Evaluation Methods

There are various testing methods for correctly evaluating our method. We have considered the option to use already tagged photos in Flickr. In this approach we selected existing photos from Flickr, took a subset of their corresponding tags to be used as initial tags, and then compared the output of the system with original tags of each photo. For statistical analysis, we calculated precision and recall values. We compared original tag list with the annotation suggestions

[2] Flickr API, http://www.flickr.com/services/api/
[3] Flickrj, http://flickrj.sourceforge.net/

we gathered by using tag frequency, color features, and SIFT similarities. Tag frequency results can be used as a base line for comparing our results with only textual tag suggestion methods.

The evaluation metric accuracy is computed as the ratio of correctly suggested tags to the total number of suggested tags. Definition of accuracy A and overall accuracy A_{Avg} are formulated in Eqn 7, where ST is the suggested tags and T is the original tags.

$$A(I_i) = \frac{\mid T_i \cap ST(I_i) \mid}{\mid ST(I_i) \mid} \tag{6}$$

$$A_{Avg} = \frac{\sum_{i=1}^{n} A(I_i)}{n} \tag{7}$$

4.4 Experiment: Suggest-All-Tags

In our analysis, our first experiment was to *Suggest-All-Tags*, where we tried to suggest all the original tags of a photo by suggesting the same number of tags to the user. Accuracy is calculated among all original tags. For instance, if a Flickr photo in our test set has originally 22 tags, and we select 2 of them as initial tags, we suggest 20 tags to the user and try to make them match with the original tag list.

Figure 3 displays several different results in the form of recall vs. precision graphs we have obtained in our experiments. From these graphs we can see the change of performance of our visual similarities, namely color histograms and SIFT descriptors, as opposed to the performance of purely text-based tag frequency approach.

For photos of (a), (b), (c) and (d) results of color features produce relatively higher performance. As it can be visually observed, colors of photos (a), (b) and (d) are very significant and make them easier to be matched to other related photos. Colors of photo (c) are not as discriminant as the other three photos and thus the performance of frequency method is closer to color histogram method.

For photos of (g) and (h) results of all three approaches are approximately close to each other and their performances are low. SIFT features are better for scenes with specific objects. However, these photos are cluttered with objects; this explains the low performance. Moreover, as these photos do not contain distinctive colors and good illumination, color features also show low performance.

4.5 Experiment: Suggest-Top-5

Aim of the *Suggest-Top-5* experiment is to retrieve the relevant tags in the top 5 suggestions, which are the most important ones for the users.

Figure 4 shows the accuracies for 15 of the photos in the set of selected Flickr photos. These results represent more or less the overall results. We can say that our method slightly increases the results in most of the instances, has a better result in some of them (d, k, l, n and o), and results are not very good in a small number of photos (f and g).

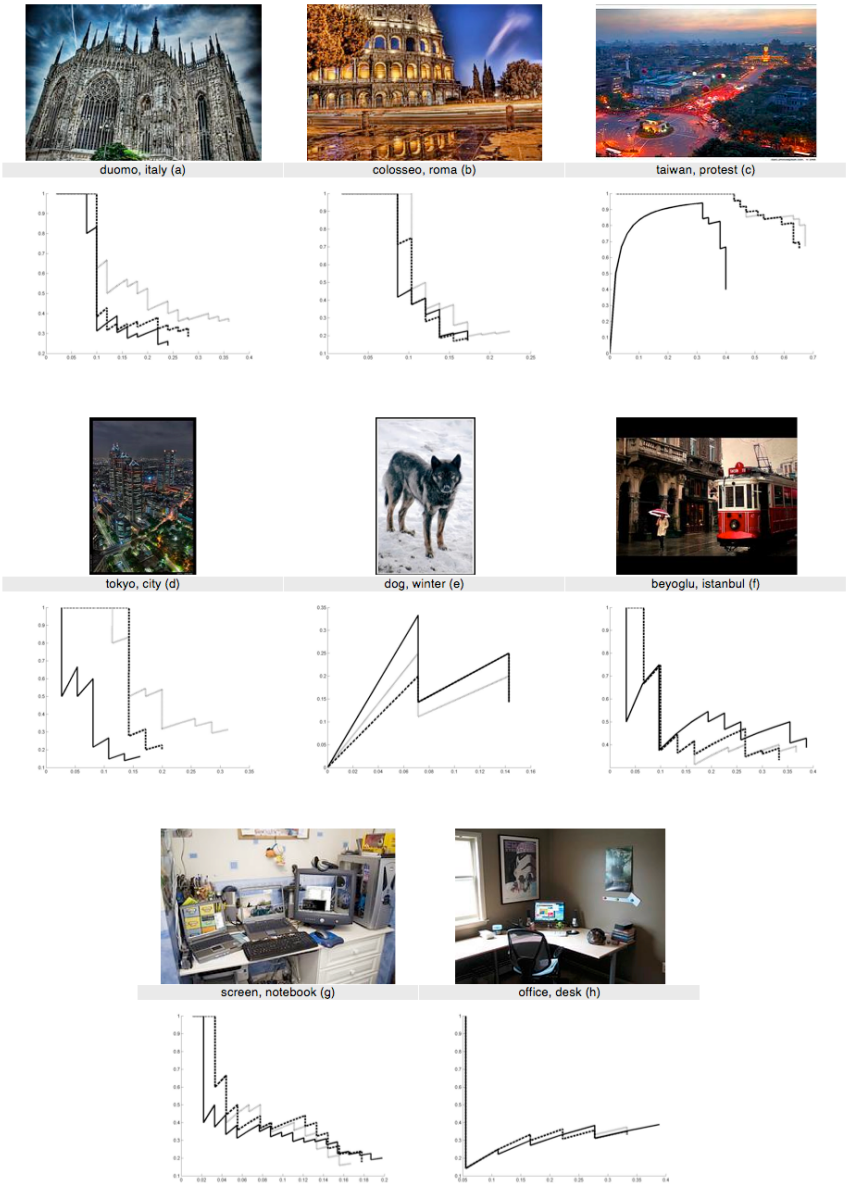

Fig. 3. Recall vs. Precision plots, their corresponding test photos together with their initial tags for Suggest-All-Tags experiment. Here, plots of color histogram method are drawn in gray, plots of SIFT similarity are drawn with a straight line, and plots of tag frequency are drawn with a dashed line. For the first four inputs (a-d), color similarity gives higher performance. Inputs (e) and (f) are exemplify high SIFT method performance. Remaining two inputs are examples for approximately close results from all three methods.

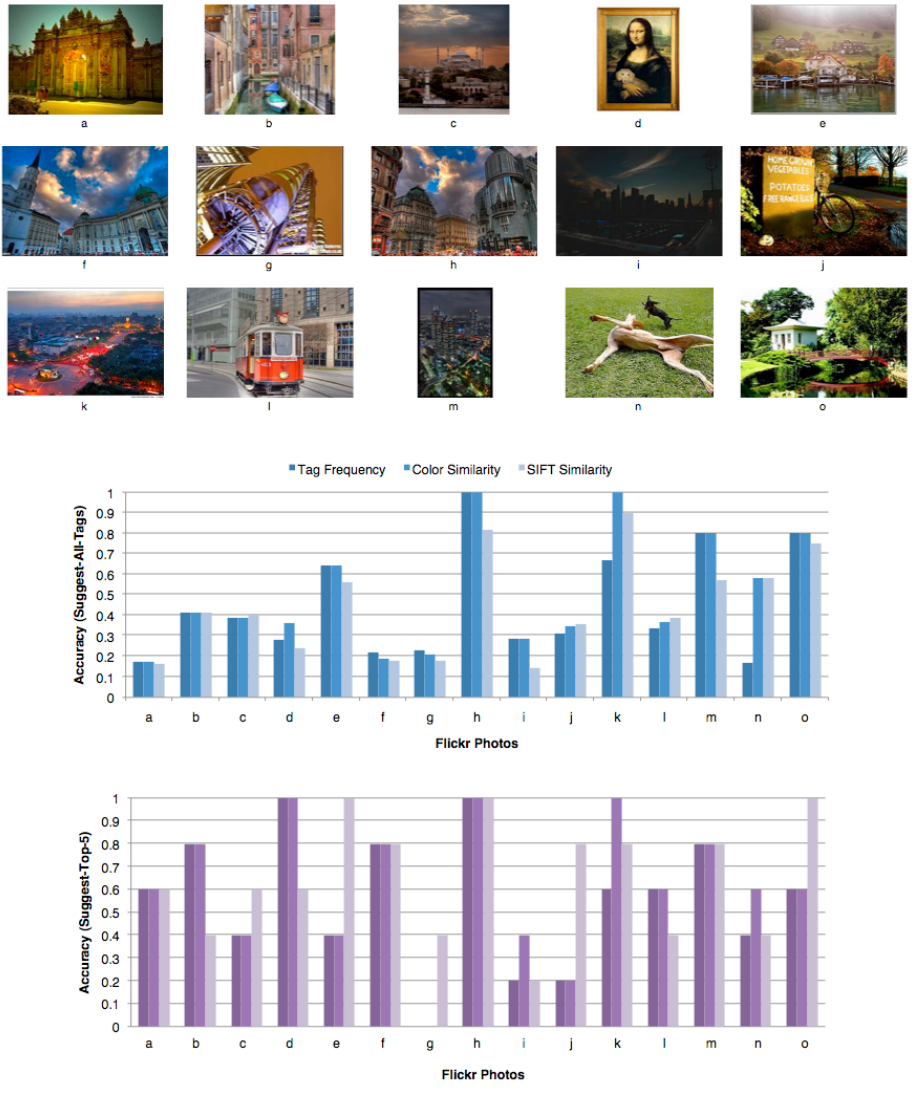

Fig. 4. Accuracy of all three methods tested on 15 given photos for Suggest-All-Tags and Suggest-Top-5 experiments. For both graphs, bars with lightest shade represent accuracy of SIFT similarity, bars with normal shade represent accuracy of Color similarity and bars with dark shade represent accuracy of tag frequency approach.

Table 2 represents the overall accuracies for both experiments. Results for color similarity are best among the three approaches. Suggest-Top-5 experiment has given significantly higher results in general because suggesting too many tags always reduces statistical performance results. Lowest accuracy is achieved when SIFT similarities are used.

Table 2. Avarage accuracy values of all three approaches for both Suggest-all-tags and Suggest-Top-5 experiments

Method	Accuracy for Suggest-all-tags	Accuracy for Suggest-Top-5
Tag Frequency	29%	52%
Color Similarity	31%	53%
SIFT Similarity	27%	46%

5 Discussion

Tagging is a very user-dependent process and validity checking of tags is difficult. Since we need to statistically evaluate tags suggested by our system, we have to define a basis for 'correct tags'. Normally, accepting user specified tags in Flickr as ground truth would be a reasonable approach. However, from our observations, we noticed that users do not properly tag their photos, and most of the tags are generally irrelevant to the image content. They have the tendency to add many commonly-used tags to make their photos popular.

Moreover, it is actually not truthful to state that, even for cases when original tags obtained from Flickr do not contain irrelevant tags, direct comparisons between a ground truth set and recommended tag set would give fully reliable results. Because for such systems, knowing that a certain tag has not been used by a user does not mean, that tag is irrelevant for that image. Different people may notice different aspects of a photo and it is not possible to represent all aspects of an image in words.

Due to these reasons, we here by claim that the statistical low performance we have observed in our experiments are not necessarily reflecting the truth. We have encountered numerous examples where our system suggested proper tags even though statistical results claimed them to be poor.

6 Conclusions and Future Work

In this paper, we proposed an automatic tag suggestion method which expands the tags of a photo by using both textual and visual information. Our motivation is to provide a system that enhances upload capabilities of a photo-sharing website so that users will easily select meaningful tags for their photos. The significance of our work is the approach of including visual information to the suggestion process. We have evaluated our system and showed that with current technologies, integrating visual features to automated systems do not add an unbearable overhead.

Automatic tagging systems have the potantial to be improved in many ways. As a future study, target photo can be examined in order to decide whether color similarity or keypoint similarity approach gives better annotation suggestions. Effective use of invariant keypoints in photos will enable our system to identify

human-made objects and logos in a photo. Furthermore, developments in computer vision techniques should improve the performance of such systems where visual similarity is involved.

For improving our statistical results, as a future work, a user study can be prepared where a photo and a list of initial tags are given, and the users choose the other relevant tags. Results of such a study can be used as the ground truth for evaluating annotation suggestion methods and would provide more reliable results.

Acknowledgments

This research is partially supported by TÜBİTAK Career grant number 104E065 and grant number 104E077.

References

1. Naaman, M., Yeh, R.B., Garcia-Molina, H., Paepcke, A.: Leveraging context to resolve identity in photo albums. In: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2005, Denver, CO, USA, June 07 - 11, pp. 178–187. ACM, New York (2005)
2. Yan, R., Natsev, A., Campbell, M.: An efficient Manual Image Annotation Approach based on Tagging and Browsing. In: Proceedings of ACM International Multimedia Conference (2007)
3. Wenyin, L., Dumais, S., Sun, Y., Zhang, H., Czerwinski, M., Field, B.: Semi-automatic image annotation. In: Proc. of Interact: Conference on HCI, pp. 326–333 (July 2001)
4. Wenyin, L., Sun, Y., Zhang, H.: MiAlbum - a system for home photo managemet using the semi-automatic image annotation approach. In: Proceedings of the Eighth ACM international Conference on Multimedia, MULTIMEDIA 2000, Marina del Rey, California, United States, pp. 479–480. ACM, New York (2000)
5. Suh, B., Bederson, B.B.: Semi-Automatic Image Annotation Using Event and Torso Identification. Tech Report HCIL-2004-15, Computer Science Department, University of Maryland, College Park, MD
6. Suh, B., Bederson, B.B.: Semi-automatic photo annotation strategies using event based clustering and clothing based person recognition. Interact. Comput. 19(4), 524–544 (2007)
7. Marlow, C., Naaman, M., Boyd, D., Davis, M.: HT06, tagging paper, taxonomy, Flickr, academic article, to read. In: Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, Odense, Denmark, August 22-25 (2006)
8. Sigurbjrnsson, B., van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: Proceeding of the 17th international Conference on World Wide Web, WWW 2008, Beijing, China, April 21 - 25, pp. 327–336. ACM, New York (2008)
9. Elliott, B., Ozsoyoglu, Z.M.: A comparison of methods for semantic photo annotation suggestion. In: 22nd International International Symposium on Computer and Information Sciences, 2007. ISCIS 2007, November 7-9, pp. 1–6 (2007)

10. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2) (2004)
11. Quack, T., Leibe, B., Gool, L.V.: World-scale Mining of Objects and Events from Community Photo Collections. In: CIVR 2008, Niagara Falls, Canada, July 7-9 (2008)
12. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content based image retrieval at the end of the early years. IEEE Trans. on Pattern Analysis and Machine Intelligence 22(12) (2000)