Multimed Tools Appl (2010) 49:81–99
DOI 10.1007/s11042-009-0394-5

Automatic tag expansion using visual similarity for photo sharing websites

Sare Gul Sevil · Onur Kucuktunc · Pinar Duygulu · Fazli Can

Published online: 14 October 2009

Abstract In this paper we present an automatic photo tag expansion method designed for photo sharing websites. The purpose of the method is to suggest tags that are relevant to the visual content of a given photo at upload time. Both textual and visual cues are used in the process of tag expansion. When a photo is to be uploaded, the system asks for a couple of initial tags from the user. The initial tags are used to retrieve relevant photos together with their tags. These photos are assumed to be potentially content related to the uploaded target photo. The tag sets of the relevant photos are used to form the candidate tag list, and visual similarities between the target photo and relevant photos are used to give weights to these candidate tags. Tags with the highest weights are suggested to the user. The method is applied on Flickr (http://www.flickr.com). Results show that including visual information in the process of photo tagging increases accuracy with respect to text-based methods.

Keywords Tagging · Photo-annotation · Visual similarity · Folksonomy · Flickr

1 Introduction

Throughout the last decade, there has been a significant increase in the number of internet users contributing to shared media available on the Web. With the help of

S. G. Sevil (✉) · O. Kucuktunc · P. Duygulu · F. Can
Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey
e-mail: sareg@cs.bilkent.edu.tr

O. Kucuktunc
e-mail: onurk@cs.bilkent.edu.tr

P. Duygulu
e-mail: duygulu@cs.bilkent.edu.tr

F. Can
e-mail: canf@cs.bilkent.edu.tr

popular social media-sharing websites and the new trend of being a part of these sites, large amounts of user-generated data in various forms have become available.

Because users tend to share photos more than any other media type, thousands of photos become available everyday. As a result, managing these massive numbers of photos for efficient and effective browsing/searching operations has become a challenging task.

To properly organize all kinds of media data, collaborative tagging methods have been used for many years where descriptive words, *tags*, are assigned to data for effective text-based search and retrieval techniques. Many photo-sharing websites, including Flickr, use this approach by giving their users the option of tagging. Unfortunately photo tagging is a very subjective and time-consuming task, especially when content descriptive tags are required to provide better navigation and searching capabilities. Studies on Flickr have shown that, although some photos contain more than 50 tags, photos with one to three tags cover more than 60% of all photos, [29]. Therefore, a tag suggestion system integrated to photo-sharing websites may be beneficial in encouraging users to add sufficient number of eligible content-related tags for their photos.

In this paper, we present Tag Suggestr, an automatic tag expansion algorithm that suggests additional tags for a given photo with two-three initial tags. The algorithm is designed for photo sharing websites where it can be used to assist users in proper tagging at upload time. The method utilizes the visual content of the target (uploaded) photo as well as the textual information obtained from the provided initial tags to form and weight a set of candidate tags. Tags with higher weights are then suggested for the target photo.

The advantage of Tag Suggestr comes from its usage of visual information during the process of tag suggestion. The purpose is to use the visual content of both the target image and the other images available within the system (photo sharing website) to be able to suggest tags that are related to the visual content of the image. This would be beneficial for navigation and searching capabilities of the system itself since content related tags would be better in identifying photos.

The goal of this approach is not to give users a complete set of tags that could be directly used but to give a list of, possibly, incomplete set of tags that would help or guide the users to tag in accordance with the image content. The most important characteristic of these suggested tags is their generality, in the sense that they do not carry information that would only be known by the photographer. Suggested tags aim to describe photos by observable concepts.

2 Related work

Social tagging of media has brought a requirement for analysis and management of large amounts of tags. This challenge is addressed in several recent studies on non-visual media resources as for del.icio.us,[1] last.fm,[2] etc. [4, 8, 22, 37].

[1] http://del.icio.us

[2] http://last.fm

Flickr, being among the most widely used social media sites, is a rich source for visual data. Due to the limitations of Content Based Image Retrieval (CBIR) methods [28, 30], other than some recent efforts for web-scale retrieval [10], accessing the content through tags is still the most commonly used way. However, since the tags are subjective and limited, automatic generation of tags is required.

As extensively analysed in [19], Flickr has some differences compared to other social media resources and therefore requires different strategies for automatic tagging. Images are most commonly tagged only by their owners, therefore tags usually represent personal aspects, and come in limited numbers. Therefore, mining the tag usage from group profiles is not in consideration, and usage of a single user's profile is not sufficient to describe the visual content required for better access, since subjectivity limits the vocabulary.

Recent studies in automatic annotation of visual data is promising [1–3, 5–7, 9, 12, 14, 20, 23, 24, 26, 35, 36]. However, most of these methods usually work on small amount of data compared to the resources on the web, and usually only a few keywords are assigned to images which may not be sufficient to describe the rich visual content of the web images.

In some approaches proposed for web images, only text-based methods are used for generation of tags. In [29], co-occurrence information for tags, obtained from a large pool of Flickr images, are used to recommend additional tags to a photo with initial user-defined tags. Since only text cues are used for generation of new tags without considering visual information, photos having same initial tags will always be tagged with the same set of new tags [15].

Recently, methods that combine textual and visual techniques have been proposed. In [16], at the first step, images are classified into a set of concepts by a multi-class SVM. At the next step, tags of visually similar photos are propagated. Since the method requires trained concepts, it can only work on specific catagories and is limited.

The work presented in [27] collects geo-tagged images from Flickr and clusters them using textual, visual and spatial cues. Using frequent itemset mining techniques, relevant word combinations are found for each cluster which are then also used to link the clusters to Wikipedia entries.

In some studies tagging by multiple users is simulated through finding similar images that can be considered as the same resource, and then learning the tagging behaviour from the tags associated with this set. The common approach is, over a large number of pre-collected pool of images, to find the neighbor images which are most visually similar to a target image, and then ranking the tags inside this set either using only frequency of tags or by incorporating visual similarities [15, 31–33].

In these approaches, visual similarities are considered as a first step to collect a set of candidate images from which the tags are used. However, this process requires the construction of a large pool of fixed images (usually in the order of millions), and computation of image similarities in such a large collection to obtain a smaller candidate set.

This cost can be overcome with the help of minimal number of textual cues provided during the query time. A few initial tags allow to collect a set of candidate images which are likely to be relevant to the target image. Visual similarities can then be further considered in order to prune this set.

In [18], given co-occurrence information of tags, for each pair including the user-defined-tag a classifier is trained. The degree of memberships of the images for a list of classifiers are then used to find the relevance of tags for recommendation. This method is very costly since a separate classifier must be built for each pair in a pre-collected co-occurrence list.

In [34], starting with a textual query, among semantically similar images, content based image search is used to find images which are also visually similar. This set is clustered, and related words for each cluster are extracted to be used as recommended tags. Since the images are pruned, and then a cluster-based assignment is performed, there is a possibility to loose some important tags.

3 Tag Suggestr

Tag Suggestr is a system designed to serve as an interface between users and photo-sharing websites during photo upload. It is applicable to all photo-sharing websites with tagging capability. In this work, we have chosen to carry out our implementations and experiments on Flickr.

When a user is uploading a photo, the method requires the user to provide a couple of initial tags that generally describe the image. These initial tags are used to retrieve a set of *relevant photos* from Flickr. Recommended tags are chosen among the distinct tags that come along with the set of relevant photos. While recommending tags, visual similarities between related photos and the photo to be uploaded are taken into account. The details of the system are described in the following subsections.

3.1 Method

Figure 1 visually describes the proposed method which can be summarized with the following steps:

Step 0 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.

Step 1 Connect to Flickr server and fetch the first m relevant photos $I_R = \{I_1, ..., I_m\}$ and their corresponding tags $T(I_i)$. Each *relevant* photo must contain the given initial tags T_{init} as a subset of $T(I_i)$.

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

Step 2 Let $T_{unique} = \{t_1, t_2, ...\}$ be the unique set of tags of all relevant photos. By subtracting the set of stopwords $T_{stopwords}$ from T_{unique}, get the candidate tag list $T_{candidate}$, which contains n distinct candidate tags.

$$T_{unique} - T_{stopwords} = T_{candidate} \tag{2}$$

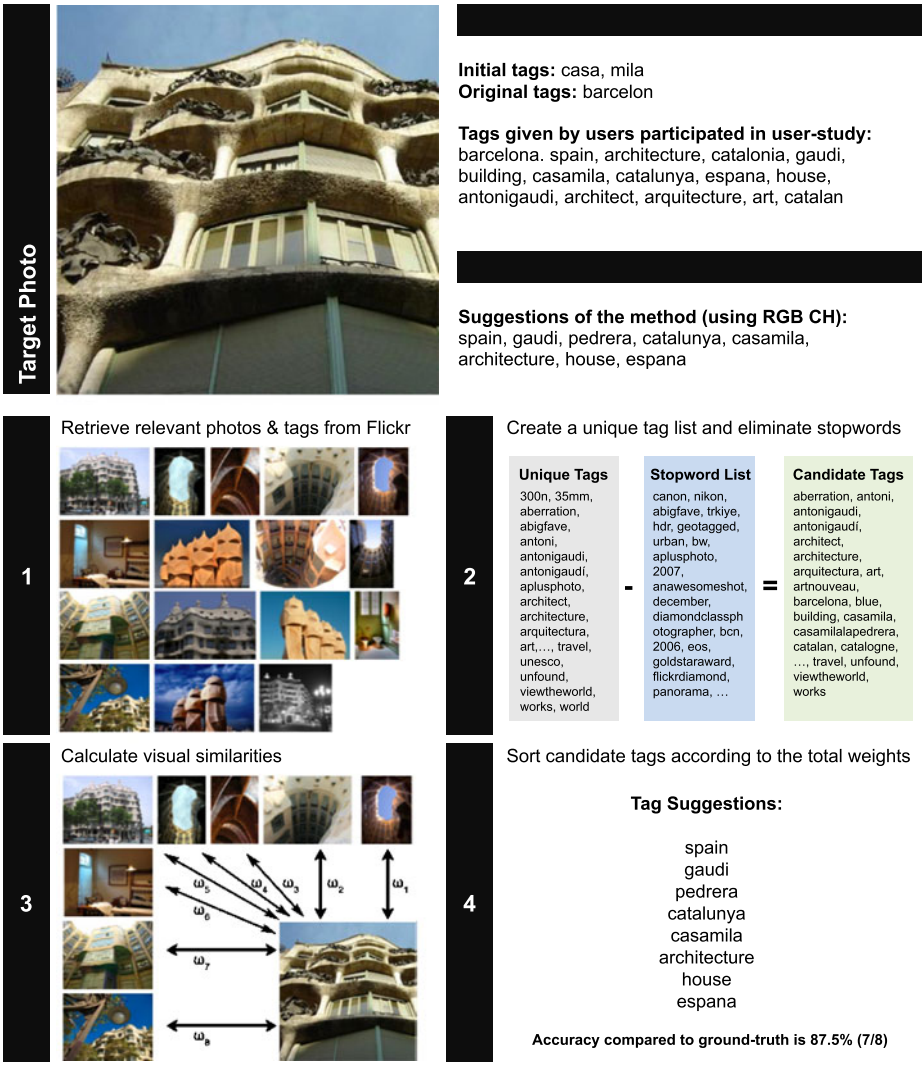

Fig. 1 Overview of the photo tag expansion method. For a given target photo and set of initial tags, the method retrieves relevant photos and their corresponding tags from Flickr (Step 1); forms a candidate tag list by eliminating stopwords from the unique tag list of all relevant photo tags (Step 2); computes visual similarities between target photo and relevant photos, then assigns weights to candidate tags using these similarities (Step 3); finally suggests tags according to their weights (Step 4)

Step 3 Extract visual features f_{I_t} for the target photo, and f_{I_i} for all relevant photos. Then, find the weight ω_i representing the visual similarity between the target photo and the ith relevant photo I_i.

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

$$\boldsymbol{\omega} = \begin{bmatrix} \omega_1 & \omega_2 & \cdots & \omega_m \end{bmatrix} \tag{4}$$

Generate a binary $m \times n$ matrix C, (where n is number of candidate tags, m is the number of relevant photos). Set C_{ij}, if photo I_i contains tag t_j.

$$C_{ij} = 1 \Leftrightarrow t_j \in T(I_i) \tag{5}$$

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 follows:

$$W = \boldsymbol{\omega} * C = \begin{bmatrix} \omega_1 & \omega_2 & \cdots & \omega_m \end{bmatrix} * \begin{array}{c} \begin{matrix} t_1 & t_2 & t_3 & \cdots & t_n \end{matrix} \\ \begin{bmatrix} 1 & 0 & 1 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \end{array} \begin{matrix} I_1 \\ I_2 \\ \vdots \\ I_m \end{matrix} \tag{6}$$

Step 4 Suggest tags in $T_{candidate}$ according to their total weights W.

3.2 Visual features

We have selected six features for computing visual similarities in our method. Descriptions of the features and similarity metrics are as follows.

RGB Color Histogram RGB color space is quantized into 27 equal subspaces; three bins per band. Visual similarity is defined as one minus the Euclidean distance between two normalized color histograms.

SIFT Descriptors The SIFT operator [17] is used to extract interest points. From these interest points, similarities between image pairs are calculated by using the matching algorithm provided by Lowe [17]. Then, *number of matching points* between two images is used as a similarity measure.

MPEG 7 Features MPEG-7 is an ISO/IEC standard that provides a set of multimedia content descriptors [21]. It was designed to functionally represent information about multimedia data for efficient searching in various applications.A set of visual features defined in MPEG-7 standards is selected for calculating the image similarities:

Color Layout Descriptor (*CLD*) captures the layout information of color feature. Because of its high retrieval efficiency and small computational costs, CLD is preferred in image and sequence matching and sketch queries.

Color Structure Descriptor (*CSD*) holds both the color content (like a color histogram) and also the structure of this content. CSD provides a better retrieval performance on natural images compared to ordinary color histograms.

Homogenous Texture Descriptor (*HTD*) provides a precise quantitative description of a texture that can be used for accurate search and retrieval. The computation of this descriptor is based on filtering using scale and orientation selective kernels.

Edge Histogram Descriptor (*EHD*) represents the spatial distribution of four directional edges and one non-directional edge. It provides better performances on image matching with non-uniform edge distribution.

An MPEG-7 feature extraction library adapted from MPEG-7 XM (eXperimentation Model) reference software is used for extracting MPEG-7 visual features [25]. l_1-norm is used as the similarity measure.

3.3 Stopword list

As explained in the previous sections, Tag Suggestr recommends tags by giving weights to candidate tags that are formed from the tag sets of relevant photos. However, these candidate tag sets usually contain many tags that are not related to the image content. For example, users interested in photography tend to flag their photos with tags describing different concepts such as camera characteristics. Also, some tags are specific to the owner of the photograph and therefore subjective. We find it appropriate to remove such tags by forming a *stopword list*, since our objective in this work is to suggest photo content related tags.

Figure 2 shows tag frequency distribution of all 25,484 candidate tags used throughout the experiments. From this distribution, we have observed the most frequent tags which appear in the top 10% of the entire data cover most of the non-image-content-related and these tags can be grouped in a stopword list. In contrast to the stopword lists used in web retrieval applications, our most frequently used tags do not include conjunction words. As there are many conjunction words, datasets

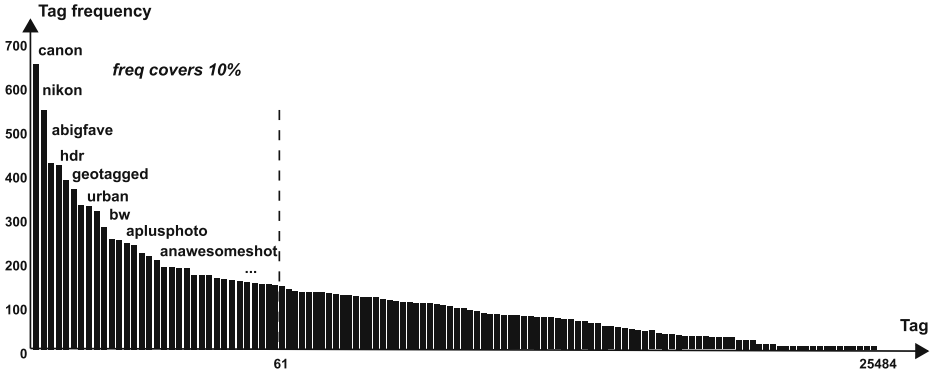

Fig. 2 Tag frequency graph of all candidate tags in the dataset. Top 10% of the most frequent tags are considered as *stopwords*

including these words have higher frequency-cut off percentages. For example, Lazarinis selects the most frequent 99 words as stopwords that covers 44.16% of the total lemmas [13]. However, this ratio is not realistic for photo tags. Therefore we selected the most frequent 61 tags that cover 10% of the total tags.

In addition to the most frequent tags, tags related to camera brands (*fuji, olympus, sony, panasonic, lumix*), camera models (*powershot, coolpix*), lens brands (*sigma, nikkor, tamron*), and photo editing software (*photoshop, photomatix, iphotoedited*), and tags with numeric values, such as years (*2008, 2009, etc.*), lens properties (*50mm, 70–300usm, etc.*), camera models (*400d, d200, f30, etc.*), and geotag information (*geo:lat=41363...*) are also eliminated when stopwords are removed from candidate tags.

4 Experimental evaluation

In our experiments, we have evaluated performance of Tag Suggestr by suggesting tags to selected target photos using six different visual features defined in Section 3.2. As in [8, 11], we use the text-based approach that suggest tags based on frequency of usage as our baseline. In the following subsections details of the experimental evaluation are presented.

4.1 Test collection

Tag Suggestr is designed to be used when a photo is being uploaded to the system. Thus we do not work on a dataset that is a pool of images to be tagged collectively. Our test collection consists of 150 arbitrary photos that are used as target photos. These photos are chosen so that contents could be easily identified by any user. This property of the photos was crucial for our user-study that is performed to form the ground-truth (explained in further detail in the following sections). As a design choice we have retrieved 100 relevant photos per target photo thus 15,000 photos have been processed throughout our evaluations. It is important to note that the number of target photos used in these experiments do not affect the accuracy of individual suggestions; since the suggested tags are chosen amongst the tags of 100 relevant photos.

Figure 3 displays the selected targets. 49 of these photos (previously used in [11]) were chosen randomly. The remaining 101 photos were more carefully chosen and they can be grouped under five city categories.

4.2 Generating ground-truth: user-study

Currently there is no ground-truth available for evaluating the performance of a tagging system. A common approach for evaluation is to use original tags of a target photo as ground-truth. However, in Flickr, tags are generally not directly content related and do not come in sufficient numbers [29]. To overcome these issues, we perform a user-study to form an unbiased and generically formed ground-truth for performance evaluation.

Since tagging is a subjective task, without limiting the vocabulary, asking users to freely tag photos would result in widely scattered lists of tags. Therefore, in the

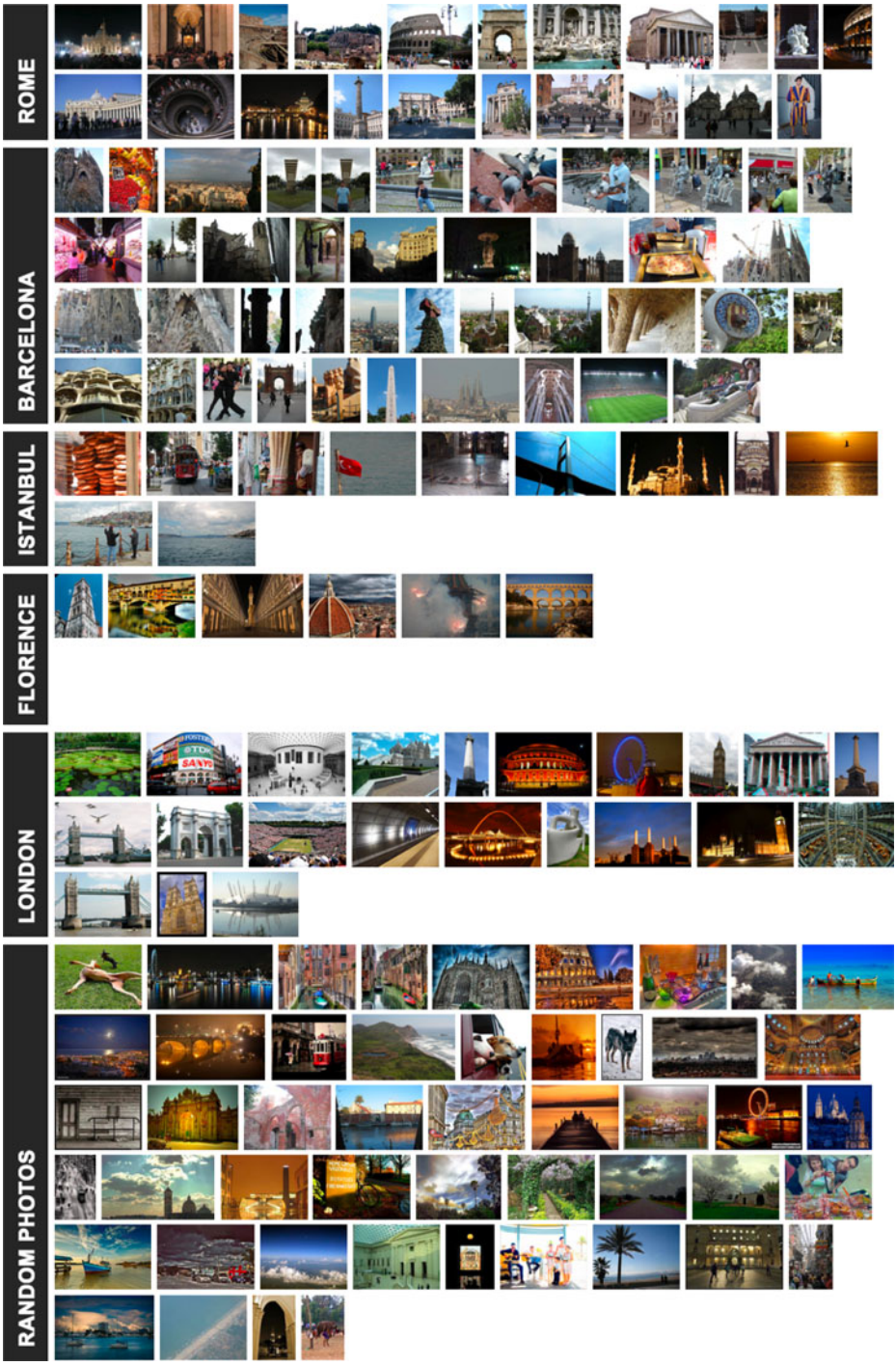

Fig. 3 Dataset of target photos selected from existing photos in Flickr. In addition to the randomly chosen photos, famous places in Rome, Barcelona, Istanbul, Florence, London are taken and initial tags are given accordingly

user-study our users are asked to select tags from the candidate tag lists, which is obtained from the tags of relevant photos.

The user-study is composed of two parts. In each part users are asked to annotate 15 different photos by selecting five tags from a given candidate tag list which is generated by our method. The two parts differed in their presentation of candidate tag lists. While in the first part candidate tags are sorted alphabetically, in the second part they are sorted with respect to their frequency of usage in relevant photos.

Users are informed that they are expanding initial tags listed below each photo. For users who are not familiar with the content of photos, links to Wikipedia entries are provided when available. Figure 4 shows a screenshot of a tagging in progress.

The user-study is completed by 80 users. Each user is given a total of 30 photos; therefore, each of the 150 target photos are annotated at least 15 times by different users. Compared to the original tags, the distribution of the tags assigned to a photo by the user-study is more homogeneous. The minimum number of tags assigned to a photo is 8, while the maximum is 40 (see Fig. 5 and Table 1).

4.3 Evaluation metrics

Performance of the method is evaluated using Precision at n, as also used in [15, 29]. We define the precision, $P@n(I_i)$, at n suggested tags for target image I_i as:

$$P@n(I_i) = \frac{|GT_i \cap ST_{n,i}|}{n} \tag{7}$$

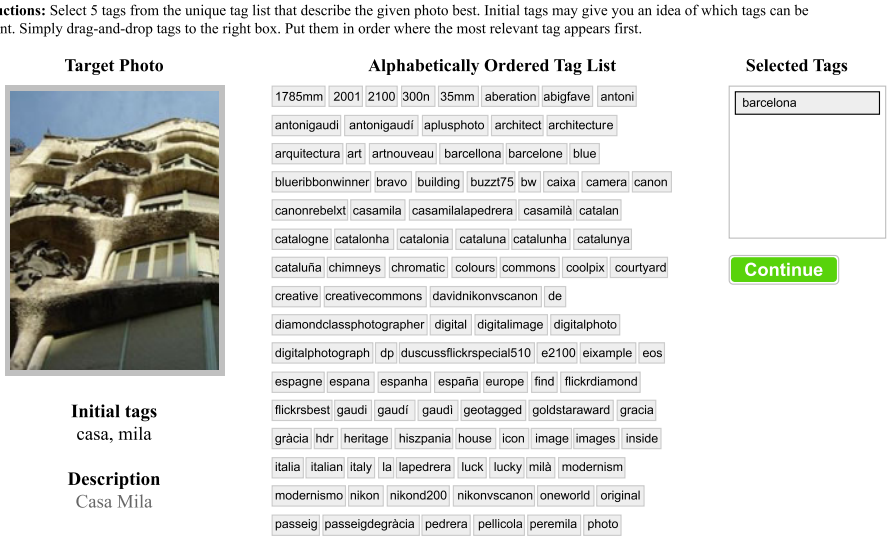

Fig. 4 Screenshot from the web-based user-study. In this experiment instance, users are asked to expand the given tags (*casa, mila*) of the photo by selecting five tags from the alphabetically sorted list. The photo itself with a link to *Casa Mila* Wikipedia entry is provided

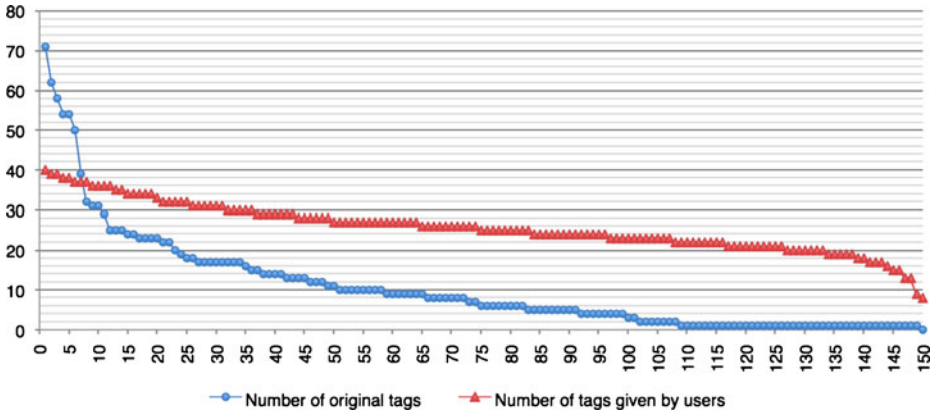

Fig. 5 Number of original tags vs. number of ground-truth tags as the result of user-study for each photo in target photo set

where GT_i is the ground-truth tags of target image I_i, $ST_{n,i}$ is the top n tags of its weighted candidate tag list (i.e. suggested tags). Three different evaluated n values are 8, 20, 25. The different n values are selected with respect to the sizes of ground-truth lists obtained from the user-study. 8 is the minimum number of tags present in ground-truth (see Fig. 5 and Table 1). 25 is the average number of tags given to a photo by the users. And finally 20 is chosen as an intermediate value. We have computed the average precision for a group of m photos as:

$$P@n = \frac{\sum_{i=1}^{m} P@n(I_i)}{m} \tag{8}$$

4.4 Results and discussion

Being one of the most intuitive approaches, tagging with the top most frequently used tags is widely used as a baseline approach for performance evaluation [15, 29]. Therefore, we also use it as a baseline method which we refer to as *frequency results* in the following. For 150 target photos under six categories (five cities and one group of random photos), Tag Suggestr is tested using six visual features. After weighting candidate tags, top n tags are chosen for suggestion. Three different values of n have been tested by checking their occurrences in corresponding ground-truth lists generated by user study.

Figure 6 shows three bar charts, one for each n value, of individual precision values for six categories and six visual features. Table 2 shows the average precision values of all 150 photos for the different visual features.

Table 1 Minimum, maximum, and average number of original and ground-truth tags

	Min	Max	Avg
Original tags	0	71	10
Tags given by users	8	40	25

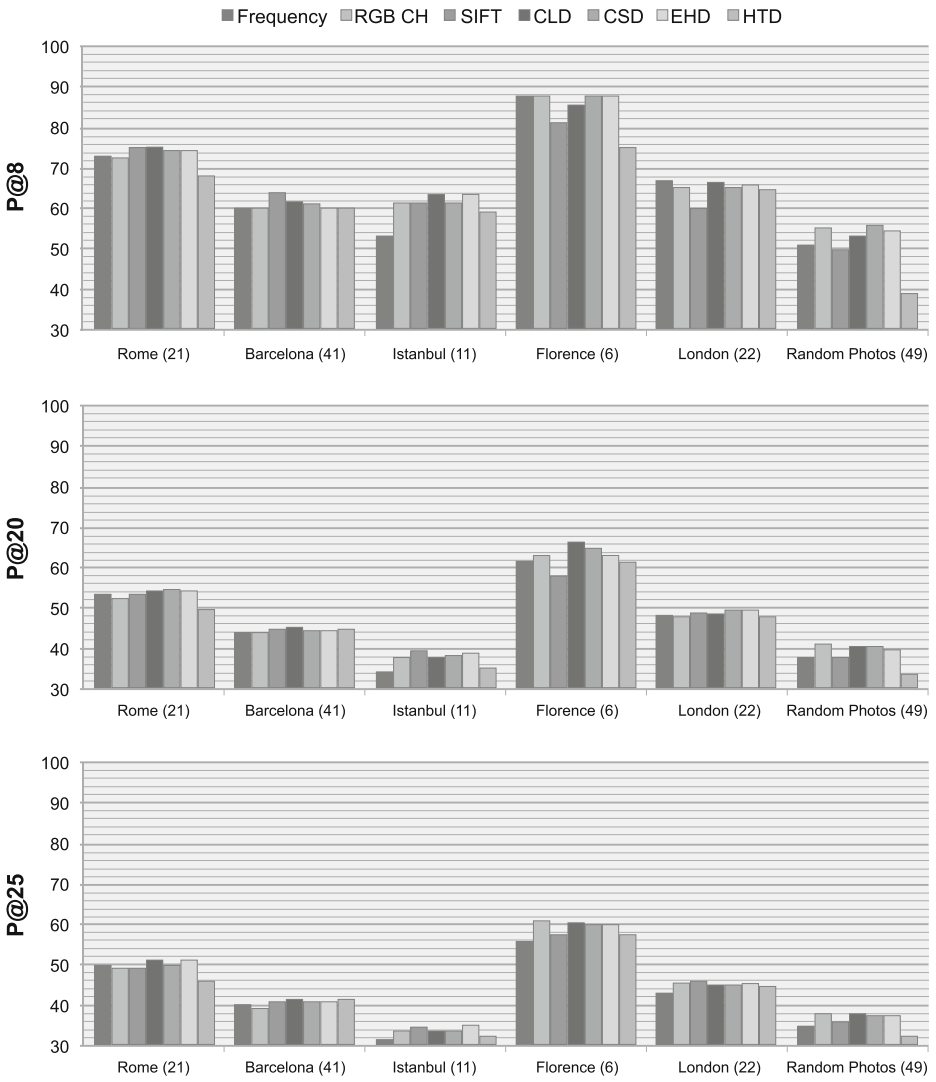

Fig. 6 Precision values of each visual feature for different photo groups. Charts (*top, middle, bottom*) show the precision values for top-8, top-20, and top-25 suggested tags, respectively. Note that the Precision axis for different evaluations does not start from 0

It can be observed that the performance of all features do not vary significantly. The reason is that the used test collection does not have a certain structure that can be easily identified by a specific low-level visual feature. The diversity of the photos produce similar performances for different tested features. In order to observe better performances, more involved visual analysis can be performed.

Selection of an appropriate test collection is important in performing a qualified evaluation process. As it can be observed from the precisions shown in Fig. 6 and Table 2, the performance is worse for photos coming from random photo group.

Table 2 Average precision of each visual feature for P@8, P@20, and P@25 using user-study tags as ground-truth

	Baseline						
	Frequency	RGB CH	SIFT	CLD	CSD	EHD	HTD
P@8	60.50	62.17	60.92	62.67	63.00	62.58	55.58
P@20	44.10	45.50	44.77	46.07	45.93	45.63	42.60
P@25	40.13	41.41	41.20	42.21	41.95	42.13	39.41

Photos within this sub-category (see Fig. 3) consist of hard-to-tag images that capture general concepts (such as animals, natural scenes, etc.) with few focused concepts to identify.

Among the remaining five sub-categories, precisions obtained from Istanbul are relatively lower. Especially for P@25, photos from Istanbul have received the worst suggestions. The reasons for this poor performance can be listed as follows: first of all, the total number of Istanbul photos in Flickr is significantly less than the number of photos available for other used categories. And we have observed that photos of Istanbul were tagged poorly. Tags mostly included too general words such as the word *Turkey* in various languages or words describing the weather. Also, relevant photos were retrieved in batches that we later realized were uploaded by the same people. Photos within these batches contained the same tags. Thus both frequency and various visual-feature based methods performed poorly. Florence, on the other

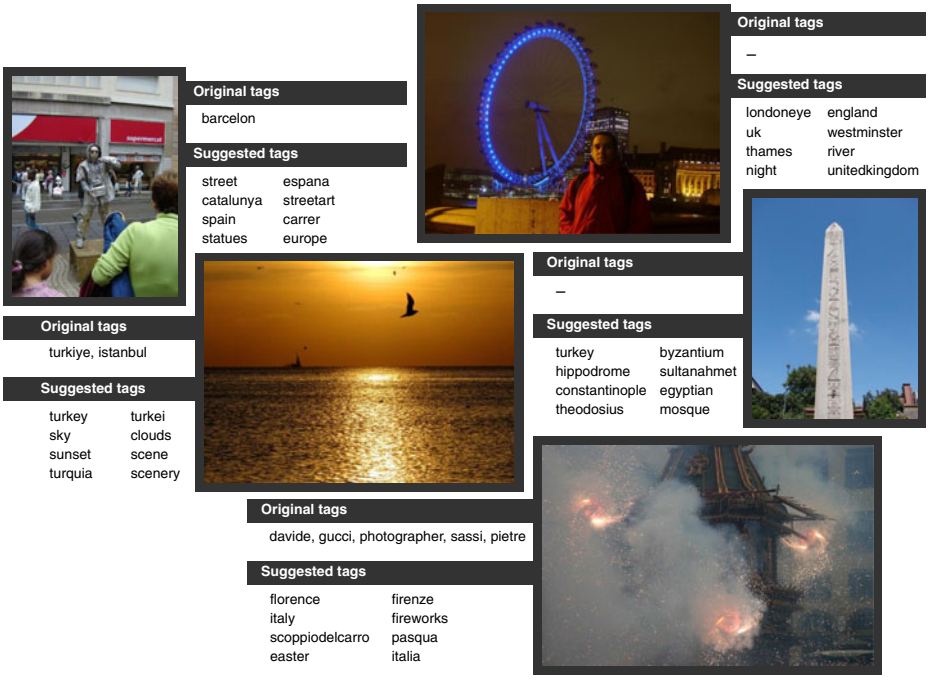

Fig. 7 Sample images from the test-collection demonstrating the differences between original and suggested tags

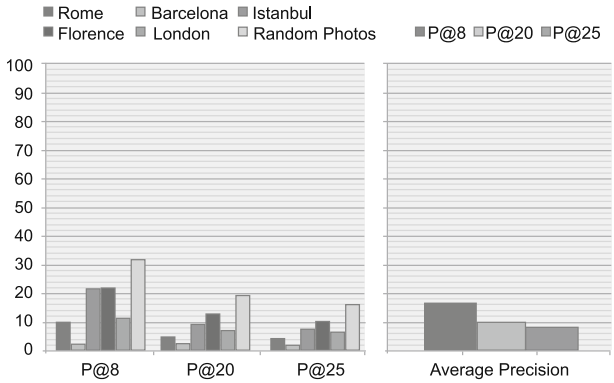

Fig. 8 Comparison of original tags with ground-truth (user-study results). Low precision values show that original tags are not eligible for proper testing purposes. Precision values of original tags for different photo groups and evaluation metrics (*left*), average precisions of original tags (*right*)

hand, has shown the highest performance among all other sub-categories. This is probably because the photos in this group were more suitable for tagging.

4.5 Analysis of original tags

As discussed previously, original tags of target images that are obtained from Flickr are limited and subjective; they are not appropriate in accessing visual content. To show this we have conducted several experiments which are explained in this section.

Figure 7 provides some visual examples of how the original tags differ from the tags suggested by our method. As it can be observed both from the graph in Fig. 5 and the images in Fig. 7, there are not enough original tags, if there are any; the original tags are not intended to specificly describe the photo content; they are subjective and possibly contain errors (e.g. misspelling of the city Barcelona as 'barcelon').

Figure 8 displays precision values of original tags when they are compared with the user-study ground-truth. On the left, we see the precision values of original tags for different photo groups and different evaluation metrics. Bars with different colors represent results for different target photo groups. It can be seen that original tags as show poor performances (maximum 30% precision) compared to our user-selected ground-truth tags in these *individual* cases. The highest performance (P@8 of original tags in random photos group which was used in [11]) can be explained by our strategy to select these photos. Due to the fact that we had evaluated our results with original tags in [11], our objective was to choose photos with higher number of tags. Since using higher number of tags improves the chance of getting high precision, original tags of the previously used photos seem to show a better performance. But when the

Table 3 Average precision of each visual feature for P@8, P@20, and P@25 using original tags as ground-truth

	Baseline						
	Frequency	RGB CH	SIFT	CLD	CSD	EHD	HTD
P@8	13.75	13.33	17.17	13.83	9.42	12.42	10.42
P@20	7.70	7.60	10.47	7.83	6.3	7.97	6.23
P@25	6.72	6.72	9.04	6.83	5.89	6.93	5.84

precisions for the complete target set are analyzed (see left chart on Fig. 8) it can be clearly seen that original tags are not eligible for proper testing purposes.

To support this point we have also conducted the all experiments using original tags as ground-truth. Table 3 shows the average precision values of these experiments; the drastic decrease of performance is observed. Figure 7 shows some examples that visually show the difference between the quality of original tags and our suggested tags.

Moreover, we have analyzed and compared tag-number distributions of original tags and the ground-truth generated by the user-study. Table 1 shows minimum, maximum and average number of tags assigned to the target photos by our ground-truth and the original tags. It can be seen that there is a larger gap between the number of tags within the original tags than the ones in ground-truth lists. The tag distribution graph in Fig. 5 supports this fact. Poor performance of the original tags discussed in the above paragraph also results from this un-even distribution of tags.

5 Conclusion and future work

In this paper we described and analyzed the automatic photo tag expansion method, TagSuggestr. We have used a large test collection from Flickr by including famous touristic places in five cities as well as a small set of randomly chosen photos. From our experiments, we have concluded that comparison of suggested tags with original tags does not reflect the true performance. As a result, a user-study was conducted in order to generate a proper set of ground-truth tags. Six visual features have been included in the analysis. Our evaluations have shown that tag expansion using mentioned visual features achieve higher precision values as opposed to text-based approach.

As a future work for better annotation suggestions, photos to be annotated can be categorized into some classes according to their visual content, and the tag expansion method with the most appropriate visual similarity can be performed. Effective use of invariant keypoints in photos will enable our system to identify human-made objects and logos/emblems in a photo. Furthermore, for systems where visual similarity is involved as ours, it is likely that better visual analysis is required to improve the performance

Acknowledgements We thank Muhammet Bastan for preparing MPEG-7 visual feature extractor, and all the users participated in the user-study. This research is partially supported by TUBITAK Career grant number 104E065.

References

1. Barnard K, Forsyth DA (2001) Learning the semantics of words and pictures. In: Proceedings of the international conference on computer vision, vol 2, pp 408–415
2. Barnard K, Duygulu P, de Freitas N, Forsyth DA, Blei D, Jordan M (2003) Matching words and pictures. J Mach Learn Res 3:1107–1135
3. Blei D, Jordan MI (2003) Modeling annotated data. In: Proceedings of 26th annual international ACM SIGIR conference, Toronto, Canada, July 28–August, pp 127–134
4. Byde A, Wan H, Cayzer S (2007) Personalized tag recommendations via tagging and content-based similarity metrics. In: Proceedings of the international conference on weblogs and social media, Boulder, CO, USA

5. Carneiro G, Vasconcelos N (2005) Formulating semantic image annotation as a supervised learning problem. In: Proceedings of IEEE conference on computer vision and pattern recognition, vol 2, pp 163–168
6. Duygulu P, Barnard K, Freitas N, Forsyth DA (2002) Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: Proceedings of 7th European conference on computer vision, vol 4, Copenhagen Denmark, 27 May–2 June, pp 97–112
7. Feng S, Manmatha R, Lavrenko V (2004) Multiple bernoulli relevance models for image and video annotation. In: Proceedings of international conference on computer vision and pattern recognition, vol 2, pp 1002–1009
8. Jaschke R, Marinho L, Hotho A, Schmidt-Thieme L, Stumme G (2008) Tag recommendations in social bookmarking systems. AI Commun 21(4):231–247
9. Jeon J, Lavrenko V, Manmatha R (2003) Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th annual international ACM SIGIR conference, Toronto, Canada, 28 July–1 August, pp 119–126
10. Jing Y, Baluja S (2008) VisualRank: applying pagerank to large-scale image search. IEEE Trans PAMI 30(11):1877–1890
11. Kucuktunc O, Sevil SG, Tosun AB, Zitouni H, Duygulu P, Can F (2008) Tag Suggestr: automatic photo tag expansion using visual information for photo sharing websites. In: Proceedings of 3rd international conference on semantic and digital media technologies (SAMT '08), Koblenz, Germany, 3–5 December 2008. Lecture notes in computer science, vol 5392/2008. Springer, Berlin, pp 63–71
12. Lavrenko V, Manmatha R, Jeon J (2003) A model for learning the semantics of pictures. In: Proceedings of 17th annual conference on neural information processing systems, vol 16, pp 553–560
13. Lazarinis F (2007) Engineering and utilizing a stopword list in Greek web retrieval. JASIST 58(11):1645–1652
14. Li J, Wang J (2003) Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans Pattern Anal Mach Intell 25(9):1075–1088
15. Li X, Snoek CGM, Worring M (2009) Learning social tag relevance by neighbor voting. IEEE Trans Multimedia (in press)
16. Lindstaedt S, Mrzinger R, Sorschag R, Pammer V, Thallinger G (2009) Automatic image annotation using visual content and folksonomies. Multimedia Tools and Applications 42(1)
17. Lowe D (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2)
18. Lux M, Marques O, Pitman A (2008) Using visual features to improve tag suggestions in image sharing sites. In: Proceedings of knowledge acquisition from the social web, Graz, Austria
19. Marlow C, Naaman M, Boyd D, Davis M (2006) HT06, tagging paper, taxonomy, Flickr, academic article, to read. In: Proceedings of the 17th conference on hypertext and hypermedia, Odense, Denmark, 22–25 August
20. Maron O, Ratan AL (1998) Multiple-Instance learning for natural scene classification. In: Proceedings of the 15th international conference on machine learning, pp 341–349
21. Martinez JM: Overview of the MPEG-7 standard. ISO/IEC JTC1/SC29/WG11 N4031 (2001)
22. Mishne G (2008) AutoTag: a collaborative approach to automated tag assignment for weblog posts. In: Proceedings of the 15th international conference on world wide web (WWW '08), Edinburgh, Scotland
23. Monay F, Gatica-Perez D (2004) PLSA-based image auto-annotation: constraining the latent space. In: Proceedings of ACM international conference on multimedia, pp 348–351
24. Mori Y, Takahashi H, Oka R (1999) Image-to-word transformation based on dividing and vector quantizing images with words. In: Proceedings of 1st int. workshop on multimedia intelligent storage and retrieval management
25. MPEG-7 XM Software (2001) Institute for integrated circuits. Technische Universität Munchen, Germany
26. Pan JY, Yang HJ, Duygulu P, Faloutsos C (2004) Automatic image captioning. In: Proceedings of the 2004 IEEE international conference on multimedia and expo, vol 3, Taipei, Taiwan, June, pp 1987–1990
27. Quack T, Leibe B, Gool LV (2008) World-scale mining of objects and events from community photo collections. In: Proceedings of ACM international conference on image and video retrieval (CIVR '08), Niagara Falls, Canada, 7–9 July
28. Rui Y, Huang T, Chang S (1999) Image retrieval: current techniques, promising directions, and open issues. J Vis Commun Image Represent 10(4):39–62

29. Sigurbjrnsson B, Van Zwol R (2008) Flickr tag recommendation based on collective knowledge. In: Proceeding of the 17th international conference on world wide web (WWW '08), Beijing, China, 21–25 April, pp 327–336
30. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
31. Torralba A, Fergus R, Freeman WT (2008) 80 million tiny images: a large dataset for nonparametric object and scene recognition. IEEE Trans PAMI 30(11):1958–1970
32. Wang G, Hoiem D, Forsyth D (2009) Building text features for object image classification. In: Proceedings of 19th international conference on pattern recognition
33. Wang C, Jing F, Zhang L, Zhang HJ (2008) Scalable search-based image annotation. Multimedia Syst 14(4):205–220
34. Wang X, Zhang L, Jing F, Ma WY (2006) AnnoSearch: image auto-annotation by search. In: Proceedings of international conference on computer vision and pattern recognition (CVPR '06), New York, USA
35. Wenyin L, Dumais S, Sun Y, Zhang H, Czerwinski M, Field B (2001) Semiautomatic image annotation. In: Proceedings of the 8th IFIP TC.13 conference on human-computer interaction (INTERACT '01), pp 326–333
36. Wenyin L, Sun Y, Zhang H (2000) 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 '00), Marina del Rey, California, United States, pp 479–480
37. Xu Z, Fu Y, Mao J, Su D (2008) Towards the semantic web: collaborative tag suggestions. In: Proceedings of third international conference on internet and web applications and services, Athens, Greece

Sare Gul Sevil received the BS degree from the Department of Computer Engineering, Bilkent University, Ankara, Turkey in 2007. She is currently pursuing the MSc degree in Bilkent University. Her research interests include automatic photo annotation.

Onur Kucuktunc received the BS degree from the Department of Computer Engineering, Bilkent University, Ankara, Turkey in 2007. He is currently pursuing the MSc degree in Bilkent University. His research interests include video processing, indexing and retrieval.

Pinar Duygulu has recently joined Department of Computer Engineering at Bilkent University. She is the co-director of RETINA Vision and Learning Group. Previously, She was a post-doctoral researcher at the Informedia project at Carnegie Mellon University. She completed my PhD, MSc and BS studies at the Department of Computer Engineering, of Middle East Technical University, Turkey. From February 2001 to May 2002, she was a visiting scholar at UC Berkeley in Computer Vision Group and Digital Library Project. During summer 2004, She was a senior researcher at CLSP Summer Workshop on Joint Visual Text Modeling at Johns Hopkins University. Her research interests include computer vision and multimedia data mining, specifically object and face recognition, semantic analysis and retrieval of large multimedia collections, historical document understanding and action recognition.

Fazli Can received the BS degree in Electrical Engineering in 1976 and the MS and PhD degrees in Computer Engineering from Middle East Technical University, Ankara, Turkey in 1979 and 1985, respectively. He is a professor of Computer Engineering at Bilkent University, Ankara, Turkey and an adjunct professor in the Department of Computer Science and Software Engineering at Miami University, Oxford, OH. In 1982 and 1983 he was at Arizona State University, Tempe, AZ and Intel Arizona for his PhD studies. Before joining Bilkent he was a tenured full professor at Miami University, and worked there between 1986 and 2005. He has published several papers in computer science conferences and journals, such as ACM Transactions on Database Systems, ACM Transactions on Information Systems, Information Processing and Management, and Journal of the American Society for Information Science and Technology. He has served on program committees of several international conferences, workshops, and NSF and TUBİTAK panels. He was one of the two co-editors of the ACM SIGIR FORUM between 1995 and 2002. He has been a recipient of the Miami University researcher of the year award, by Sigma Xi and is one of the co-founders of the Bilkent Information Retrieval Group.