Geo-based Automatic Image Annotation

Hatem Mousselly Sergieh
INSA de Lyon
7, Avenue Jean-Capelle
69621 Villeurbanne, France
hatem.mousselly-sergieh@insa-lyon.fr

Mario Döller
University of Passau
Innstrasse 43
94032 Passau,Germany
mario.doeller@uni-passau.de

Elöd Egyed-Zsigmond
INSA de Lyon
7, Avenue Jean-Capelle
69621 Villeurbanne, France
elod.egyed-zsigmond@insa-lyon.fr

Gabriele Gianini
University of Milan
via Bramante, 65
26013 Crema, Italy
gabriele.gianini@unimi.it

Harald Kosch
University of Passau
Innstrasse 43
94032 Passau,Germany
harald.kosch@uni-passau.de

Jean-Marie Pinon
INSA de Lyon
7, Avenue Jean-Capelle
69621 Villeurbanne, France
jean-marie.pinon@insa-lyon.fr

ABSTRACT

A huge number of user-tagged images are daily uploaded to the web. Recently, a growing number of those images are also geotagged. These provide new opportunities for solutions to automatically tag images so that efficient image management and retrieval can be achieved. In this paper an automatic image annotation approach is proposed. It is based on a statistical model that combines two different kinds of information: high level information represented by user tags of images captured in the same location as a new unlabeled image (input image); and low level information represented by the visual similarity between the input image and the collection of geographically similar images. To maximize the number of images that are visually similar to the input image, an iterative visual matching approach is proposed and evaluated. The results show that a significant recall improvement can be achieved with an increasing number of iterations. The quality of the recommended tags has also been evaluated and an overall good performance has been observed.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; I.4 [**Image Processing and Computer Vision**]: Applications

General Terms

Algorithms, Experimentation, Performance

ICMR '12, June 5-8, Hong Kong, China

Keywords

Image Annotation, Geotagging, Image Retrieval, Statistical Models

1. INTRODUCTION

Recent technological development has allowed an easy access to digital photography devices, such as digital cameras and smart phones. Everyday, individuals are producing large amounts of images and upload them to the web. For example, Flickr[1] hosts about 6 billion images. As a result, image organization and retrieval became more challenging and efficient solutions are required.

Research in content based image retrieval (CBIR) [6] [5] addresses the problem by identifying, combining and comparing different kinds of low level image features, such as color histogram, texture, etc. Although this alleviates the problem, CBIR alone is still unable to cope with the semantic notions of user queries. In fact, most CBIR systems follow a query by example (QBE) scheme which is still far too less spread than the conventional keyword-based image search.

To narrow the semantic gap, annotating images with keywords, known as tags is widely used. Tagging enables easier management and retrieval of images. Recently, in addition to the huge amount of user tagged images which are available on the web, the number of images which are provided with location information "geotagged" is increasing. Geotagged images contain in their Exif descriptors [16] the coordinates (latitude and longitude) of the location of their capture. Currently, Flickr[2] hosts more than 171 million geotagged items, while a statistic of the year 2007 shows that Panoramio[3] hosts 2 million geotagged images.

The high availability of user- and geotagged images provides opportunities to develop new tools for automatic image annotation. This can be done by combining the high level

[1] http://news.softpedia.com/news/Flickr-Boasts-6-Billion-Photo-Uploads-215380.shtml
[2] http://www.flickr.com/map
[3] http://blog.panoramio.com/2007/06/2-million-photos-in-panoramio.html

information represented by user tags and CBIR solutions which work on image low level features. For better understanding, consider the following scenario.
Assume that a user took a photo of the building of *Institute de France* (Fig. 4a) with his smart phone. The GPS receiver of the phone calculates the latitude and the longitude of the location and the data are stored in the Exif descriptor of the produced image. Later on, the user wants to upload this image to his favorite social website but he doesn't have time to tag it, or he doesn't even know about the place and the contents of the image. The user can send the image to a tag recommendation system which extracts the GPS coordinates of the image. Then, the system applies a geo-based search on online image databases, such as Flickr to find user-tagged images which were taken by other users in the surroundings of *Institute de France*. From the retrieved set, the system tries to find images in which the building of *Institute de France* appears. For this purpose, the system searches for images from the retrieved collection which are visually similar to the input image. Finally, the tags of the visually matching images can be analyzed to produce annotation suggestions for the input image.

To address each of the steps presented in the above image annotation scenario, a statistical model for automatic image annotation is proposed. It combines high level (user tags) and low level information extracted from community image to annotate a new image taken in the same location. To maximize the size of the candidate tag collection an iterative image matching approach is provided. It increases the number of visually similar images and, as a result, the number of the collected tags. This has a direct influence on the quality of the provided tags. First, additional candidate tags can be discovered. Second, form statistical point of view, the usage pattern of user tags can be better estimated in a larger tag collection. To reduce the runtime of the visual matching, an image clustering approach based on image low level feature similarity is introduced. Finally, different evaluation studies for estimating the effectiveness of the introduced solutions are provided.

The rest of the paper is organized as follows: in the next section related work is reviewed. In section 3 a detailed description of the annotation approach is presented. In section 4 different evaluation scenarios are provided. The paper is concluded and future work is discussed in section 5.

2. RELATED WORK

This work is related to search-based image annotation approaches. It exploits geo- and user-tagged image datasets and proposes a probabilistic model for tag recommendation. This model combines low level image features and user tags to produce word-image relevance values which can be used to rank the recommended tags. In the following, we review image annotation works which use similar techniques.

The authors in [18] introduce a hybrid probabilistic model for automatic image tagging. The model combines CPAM (Colored Pattern Appearance Model) [15] image features with textual information provided by the user as initial tags. The goal is to extend the set of provided initial tags by recommending new ones.

In [11] user-tagged community photos are exploited to provide a tag ranking approach. Hereby, the importance of the tag is determined by the number of visual neighbors of the input image which are annotated with that tag. The visual neighbors are identified based on three image features, namely, color correlgoram, color texture moment and RGB color moment.

The tag recommendation approach presented in [17] generates tag rankings by combining three kinds of tag correlations is presented in: tag co-occurrence in a large Flickr image dataset, tag correlation calculated based on visual language model (VLM), and image-tag conditioned tag correlation.

The image annotation approach proposed in [12] combines different search schemes to estimate the relation between a word and an unlabeled image. For this purpose, keyword-based image retrieval is used to calculate the likelihood of obtaining a certain image given a certain keyword. After that, the set of candidate annotations is expanded by applying a linear combination of two kinds of search-based word correlations. A statistical correlation that uses the words as input to Google image searcher and calculates "Normalized Google Distance" [4] based on the provided results. The other correlation is based on the visual consistence among the resulted images returned by the search.

A geo-based image annotation approach for landmark photos is proposed in [7]. It uses multi-level clustering approach that exploits geographical as well as visual features to identify the category of an input image and the nearest visual clusters. A caption for the input image is generated form the set of most frequently used tags in the corresponding visual cluster.

Among the above introduced works only [7] consider geo-tagged community photos. In fact, our approach share similar ideas, however, we have a broader focus than only providing captions for landmark photos. We provide a solid framework for tag recommendation that exploits visual, geographical and different tag occurrence measures to provide ranked tag suggestions, while image captions produced in [7] depend only on the usage frequency of the tags in the visual cluster.

The statistical model provided here is similar to the one proposed by [18]. However, our approach considers the geographical context of the input image and uses a different family of low level image features. Furthermore, we don't expect any further input for our approach except the input image, while initial tags are considered as perquisite to generate tag recommendations in [17]. In [18] the absence if initial tags leads to a pure CBIR search.

Finally, no one of the approaches presented in this section considers the effect of the volume of the image dataset on the quality of the suggested tags. In this paper we deal with this problem and propose an iterative visual matching solution to find additional visual correspondences for an input image. We also show how the size of the visual cluster influences the quality of the produced tags.

3. TAG RECOMMENDATION APPROACH

The proposed approach consists of three phases (see Fig.1). A *geo-based search* in which the geographical coordinates of an input image are used to find a collection of user-contributed images that are taken in the same geographical area. Hereby, the web image databases Flickr and Panoramio are used since they contain a large amount of geo- and user-tagged images and provide a freely available API. Next, a *visual matching* is applied on the retrieved image collection to identify images that are visually similar to the input im-

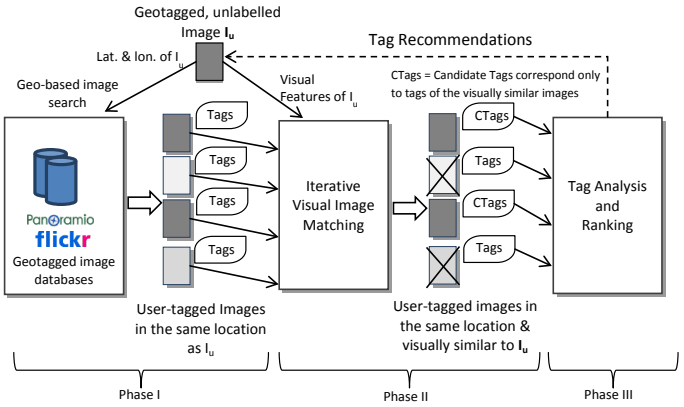


Figure 1: Three-phases automatic tag recommendation approach

age. Finally, user-tags of the visual correspondences are *analyzed* and used to generate *tag recommendations* for the input image. The output of each phase is integrated into a probabilistic model that generates scores (ranks) representing how suitable is a candidate tag for the input image.

3.1 Problem Statement

Let I_u be an unlabeled image taken at a location (lt, ln) where lt and ln represent the latitude and the longitude of the location respectively. Let $C = \{I_1, I_2, ..., I_m\}$ be a set of images taken in the same geographical area as I_u, such that $\forall I_i \in C$, $distance(I_u, I_i) < d$, where $distance(I_u, I_i)$ is the geographical distance between the locations of capture of the two images and d is a predefined distance threshold. Furthermore, suppose that members of C are annotated with words taken from the lexicon $W = \{w_1, w_2, ..., w_n\}$. Each word in W can be used at most for one time to annotate the same image, however, the same word can be used to annotate more than one image.
The introduced components and their relationships can be visualized by an undirected graph $G(V, E)$ where the vertices corresponds to elements from $V = W \cup C$ (the union of word and image sets). Edges connect words and images and are defined as $E = \{(w, I)|w \in W \wedge I \in C\}$. An edge indicates that a word has been used to annotate the linked image. The problem of recommending a tag to an input image can be expressed by finding edges between the vertex representing the input image and the set of vertices which correspond to words. Such edges are represented as dashed lines in Fig. 2. To achieve that, an edge can be drawn between I_u and a word w if there is an image $I_i \in C$ which is visually identical to I_u (i.e. a copy of I_u) and annotated with w. However, this assumption is too strict and to loosen it, we assume that an edge can be drawn between I_u and a word w if some kind of similarity relationship exists between I_u and an image I_i which is annotated with w. There are different possibilities to derive such a relationship: in this paper, we will use Bayesian methods to derive a relationship from the visual similarity between the images I_u and I_i.

3.2 Pseudo-generative Statistical Model

We frame the problem within a probabilistic model based on Bayes' Rule (a.k.a. Total Probability Law). Paradigmatically the Bayes' Rule is used in settings where, from a generative model– linking some causes (endowed with an a-priori probability) to some effects – one wants to get the probability of the effects given a cause. This generative model is typically represented by means of a tree, where

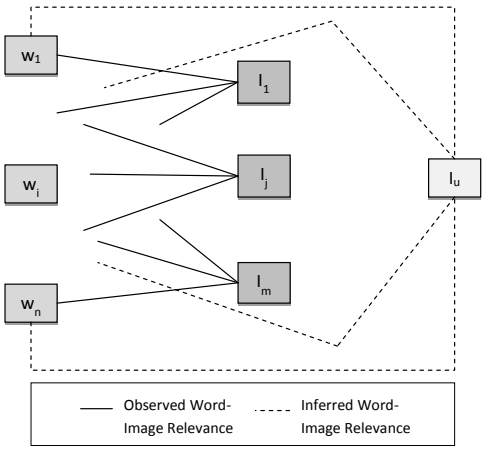


Figure 2: Graph representation of word-image relevance

the leaves correspond to the effects, (or by means of a more compact graph where the effects can be shared among different causes). Complying to this paradigm, we can identify, for sake of convenience, the input image I_u with the root of the tree, and the words w_i with the leaves, whereas the images I_j play the role of intermediate nodes, "caused" by the root and "causing" the leaves. Within this pseudo-generative model the input image I_u can be thought to "yield" each of the images $I_j \in C$, with a probability value $P(I_i|I_u)$; in turn images from C can be thought to "yield" the words w_i – which annotate them – with probability $P(w_i|I_j)$. This model in the compact version is shown in the graph of Fig. 3.

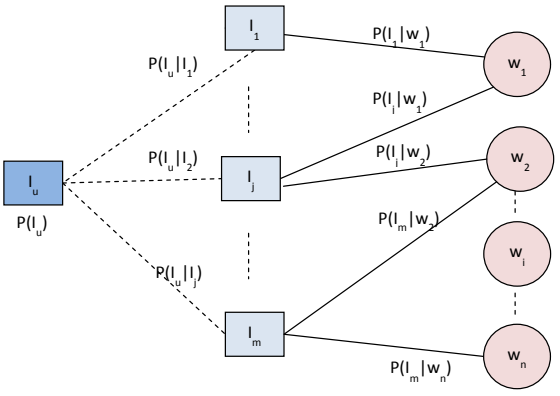


Figure 3: A Bayesian model for tag recommendation. I_u represent the input image, $I_j \in C$ is the collection of images found in the same geographical area as I_u and $w_i \in W$ are the set of tags associated with members of C

Let us assume that the values of the conditional probabilities $P(I_j|I_u)$ and $P(w_i|I_j)$ are available, then we can define the strength of the relationship of I_u with a word w_i to be equal to the conditional probability $P(w_i|I_u)$, which, given an input image I_u, can be computed by means of the Bayes'Rule as

$$P(w_i|I_u) = \sum_{j=1}^{m} P(I_j|I_u)P(w_i|I_j)$$

This corresponds to the sum over all the paths leading from the root image I_u to the leaf word w_i, of the path's probability computed by chain product. Therefore, by construction, only words that are reachable from the input image are considered as candidate tags. Once the probabilities $P(w_i|I_u)$ have been computed for all the candidate tag words relevant to a specific input image, the probability values provide a natural ordering of the importance/appropriateness of the words to the input image, and few of them, the top ranking ones, as actual input image tags.

The assumption of the availability of the values of the conditional probabilities $P(I_j|I_u)$ and $P(w_i|I_j)$ needs further consideration: it is a strong assumption, however it can be easily relaxed. Strictly speaking neither the generative process from I_u to the I_j 's nor the generative model from the I_j's to the w_i's are known or defined precisely, hence the above conditional probabilities cannot be known exactly. However we are not interested in probabilities "per-se", but rather in probability values as indicators used eventually for ranking the different candidate tag words (for appropriateness with respect to a specific input image). Therefore even quantities proportional to (or simply monotonically dependent on) those probabilities will suite the task, because they will not change the ordering. Furthermore if the probability gaps between pairs of images ($P(I_j|I_u) - P(I'_j|I_u)$ with $I_j, I'_j \in C$) relevant to an input image are wide enough, then even slightly distorting functions or indicators correlated with the $P(.|I_u)$ (proxies of the $P(.|I_u)$) can suite the task. In an analogous way proxies of the $P(.|I_u)$ can can be used in their place if the probability gaps between pairs of candidate tag words describing a tagged image ($P(w_i|I_j) - P(w'_i|I_j)$ with $w_j, w'_j \in W(I_i)$) are wide enough . For those reasons even if the conditional probabilities $P(I_j|I_u)$ and $P(w_i|I_j)$ are not directly available to us, we will adopt the above described ranking procedure: in place of the probabilities we will use proxy quantities – respectively an image-to-image similarity measure and a word-to-image importance measure – which are introduced in the next subsection.

3.2.1 Image-to-Image Similarity

The term $P(I_i|I_u)$ represents the probability of generating the image I_i from the input image I_u which was taken in the same geographical neighborhood. There exist different kinds of information that can be exploited and combined to calculate this probability. For example, some of this information can be extracted from user profile similarities and other from the geographical context of the image pair, such as the geographical distance. In this paper, we only consider image visual similarity based on low-level image features to estimate this probability.

Visual Similarity

Images which are taken in a certain location vary according to the camera perspective, zoom, light conditions, etc. Correspondingly, finding visually similar images requires matching the images based on low level features that are robust against changes in the scale, rotation, skew and illumination. Recently, various approaches for image matching based on finding correspondences among distinctive points, called interest points, have shown a great success in wide range of applications. The general approach can be described in a three step process. First, a *detector* is used to identify *interest points* in the image, such as corners or blobs. Next, every interest point is represented by a distinctive feature vector known as *descriptor*. Finally, correspondences are determined based on the distance (Euclidean distance or Mahalanobis distance) between the two descriptors. Among the different image matching algorithms that follow this approach we selected SURF (Speeded Up Robust Features) [2]. The SURF detector and descriptor are scale and rotation invariant. A special characteristic of SURF is also that it uses a small sized descriptor. This allows fast matching and makes the algorithm suitable for online applications. To achieve this, SURF uses integral images to reduce the computation required by a 'Hessian' detector. Additionally, SURF descriptor is built from Haar-wavelet responses within the interest point neighborhood and has a size of only 64 dimensions. In [8] it has been shown that SURF outperforms other algorithms from the same family, namely SIFT [13] and PCA-SIFT [9]. The evaluation shows that SURF is faster than the other approaches and provides matching results comparable to that of SIFT.
To estimate the term $P(I_i|I_u)$, the visual similarity between the two images is calculated. For this purpose, the SURF descriptor of each image is extracted and common interest points are determined. The similarity of the image pair can be calculated by computing the Dice's (intersection over union) similarity coefficient, which compares the number of common elements with the total elements of two sets. The Dice's coefficient is defined as follows:

$$Sim(I_i, I_j) = \frac{2 \times |IPs(I_i) \cap IPs(I_j)|}{|IPs(I_i)| + |IPs(I_j)|} \quad (1)$$

where:
- $IPs(I_i)$: The set of interest points extracted from I_i.
- $IPs(I_j)$: The set of interest points extracted from I_j.
- $IPs(I_i) \cap IPs(I_j)$ The set of common interest points. It consists of interest point pairs which have minimum Euclidean distance between their descriptor vectors[4].

Now a proxy of the probability $P(I_i|I_u)$ can be obtained by normalizing the visual similarity between I_i and I_u according to the total similarity between I_u and all images form C.

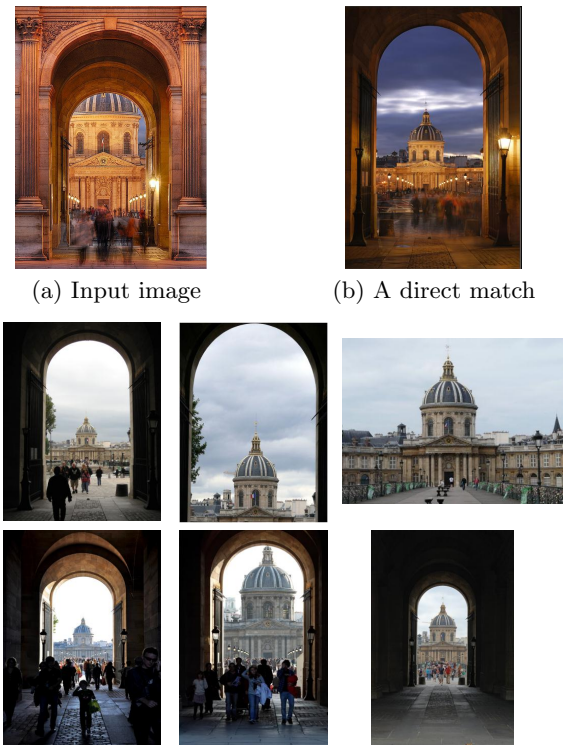
$$P(I_i|I_u) = \frac{Sim(I_i, I_u)}{\sum_{j=1}^{m} Sim(I_j, I_u)} \quad (2)$$

Iterative Image Matching

Approaches for interest point based image matching fall short of discovering similarity between two versions of the same image, for example, if the image is rotated by an angle exceeding some threshold. In our scenario, this results in loosing images which are visually similar to the input image, thus, additional candidate tags can also be missed.

The recall of image matching that is based on interest points similarities can be improved by applying the matching algorithm repetitively. The idea is very simple, given an input image, visual correspondences that are found by the matching algorithm can be used as a new input to further applications of the algorithm. This results in finding an increasing number of visually similar images, thus, the matching recall can be improved. To better understand the idea refer to Fig. 4. An input image shown in Fig. 4a is

[4]We used a Java implementation of the algorithm using the standard settings as described by [2]. The implementation is provided by Eugen Labun and is available at: http://homepages.thm.de/~elbn98/imagej-surf/



(a) Input image (b) A direct match

(c) Indirect matching images

Figure 4: Iterative interest point-based image matching

visually matched based on the SURF descriptor with images found in the same geographical location. One of the found matchings is shown in Fig. 4b. Although other images (e.g. Fig. 4c) are also visually similar to the input image they have not been discovered by the matching algorithm. For this purpose, a new visual matching is performed using the image in Fig. 4b as input. This results in finding additional matching images as can be seen in Fig. 4c. Those images are, in turn, visually similar to the initial input image. The introduced matching approach can be represented in a tree structure[5] as it is shown in Fig. 5. The root of the tree corresponds to the original input image. Visually similar images which are connected with a single edge are considered as a direct match. The iterative matching implies an increased computation which reaches its maximum when a naive method is applied in which all the matching images from a previous iteration are used as input for the next iteration. To reduce the computation cost, input of the next matching iterations have to be wisely selected. It can be noticed that visual correspondences of an input image can intersect in the sets of interest points which they share with the input image. Therefore, it can be sufficient for the next matching iteration to only consider images that match the input image according to different set of interest points. To achieve that, we propose clustering the visual correspondences of an input image based on the common interest points which they share with the input image. Consequently, only one representative image from each cluster

[5]This tree representation only holds if visual similarities among images resulting from matching the input image are ignored. Otherwise, the structure corresponds to a graph.

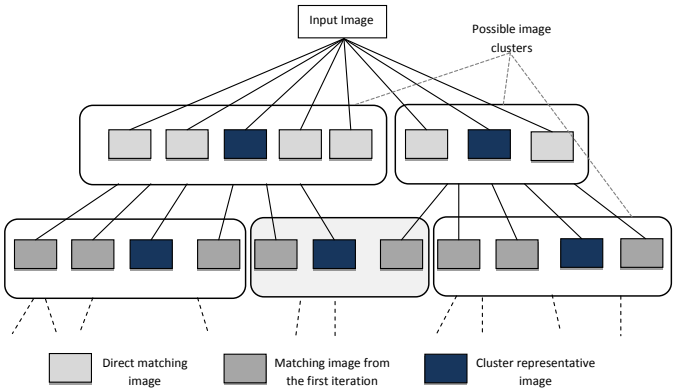


Figure 5: Tree representation of iterative image matching with clustering considerations. The rounded rectangle around an image group indicates that those images match the input image according to the same (sub)set of interest points. Therefore, only one of them, i.e., the cluster representative will be used as input for the next matching iteration.

can be used as input for the next matching iteration (Fig. 5). Creating the clusters also implies comparing image descriptors. However, the number of compared descriptors is sharply reduced because only the descriptors of the interest points that correspond to the input image are considered. In general, this number corresponds to tens of interest points and in most cases to less than ten, while the whole number of interest points extracted from an image can exceed thousands.
The effectiveness of the iterative image matching approach and the trade off between using the naive and the clustering approach is evaluated in section 4.1.

3.2.2 *Candidate Word Importance*

The term $P(w_i|I_j)$ can be interpreted as an indicator to how mutually discriminative are an image and a word. We directly derive the proxy value for this term form the model seen in Fig. 3 using a simple frequency based approach as follows:

$$P(w_i|I_j) = \frac{1}{\text{Number of words annotating } I_j} \qquad (3)$$

This measure assumes a uniform probability for all words concerning a given image. This value increases when a smaller number of tags are used to annotate the image. The value of $\sum_{j=1}^{m} P(w_i|I_j)$ can be considered as a weighted voting for the word. Each vote for a word w_i given by an image I_j is scaled by the inverse frequency of the words annotating that image.
In this paper we aim to compare the above presented measure to a TF-IDF (Term Frequency-Inverse Document Frequency) based approach. In a similar manner to [14], an adapted version of the TF-IDF measure can be used to estimate the word probability while considering a set of images annotated with it.
Since we assume that each word can be used only once to annotate a given image, one to one correspondence between an image and a document results in the same term frequency

value for all tags annotating that image, i.e., one to the total number of tags. To deal with this problem, a more realistic measure for term frequency can be defined as follows: let $R \subset C$ be the set of visual neighbors of the input image and $R_w \subset R$ the subset of visual neighbors which are annotated with w. We consider R_w as one document and calculate the term frequency of the word w as follows:

$$TF(w) = \frac{|R_w|}{|R|} \tag{4}$$

The inverse document frequency for w is calculated in a traditional way. Assume that each image in C corresponds to a document and let $C_w \subset C$ indicate the subset of images in C which are annotated with w, the inverse document frequency can be calculated as follows:

$$IDF(w) = log_2(\frac{|C|}{|C_w|}) \tag{5}$$

Now, the TF-IDF importance of a word w can be estimated as:

$$P(w_i|I_j) = \frac{TF(w).IDF(w)}{log_2|C|} \tag{6}$$

The term $log_2|C|$ is used for normalization.

3.2.3 User Related Issues

Since tags are contributed from users, the importance of a tag as a recommended annotation candidate can be determined by its usage pattern by groups of users. Generally, users tend to use the same set of words to annotate their image collections even if the images share a little or no visual similarity. This results in increasing the usage frequency of some words in a way that does not reflect its importance for automatic annotation. To address this problem, we propose scaling the importance value of a candidate word by a user's factor: U. In this respect, we propose two ways to calculate U. One way is to take into consideration the whole set of unique users who tagged images captured in a given geographical area. In this case, U can be calculated as the proportion of unique users who annotated images using some word w_i to the total number of unique users:

$$U_{\text{all users}} = \frac{\text{number of unique users using } w_i}{\text{total number of unique users}} \tag{7}$$

On the other hand, U can be estimated by only considering unique users who used w_i. In this case, U can be given as:

$$U_{\text{users of } w_i} = \frac{\text{number of unique users using } w_i}{\text{total number of occurrences of } w_i} \tag{8}$$

In the next sections we will show how the quality of the recommended tag can vary according to the provided measures for word importance as well as the user considerations.

4. EVALUATION

In this section different evaluation scenarios are presented. In the first part, we will show the effect of an iterative image matching on improving the matching recall. Hereby, we compare two variations of the algorithm: a naive and a clustering based approach. The evaluation reports the achieved precision and recall and the runtime taken by each method. The second part is concerned with evaluating the annotation performance. For this purpose, we created our ground truth and used a subset of the NUS-WIDE [3] dataset and calculated the average precision and recall.

4.1 Effectiveness of Iterative Image Matching

The ground truth for this test consists of groups of images crawled from Flickr, Google Images and the European Cities 50K dataset [1]. Each group of the datasets contains images which have the same contents but vary in scale, rotation, and illumination. Moreover, the dataset expands over images of different categories, such as outdoor, indoor, buildings, nature etc. In total, we gathered 69 such groups with average of 70 images each.

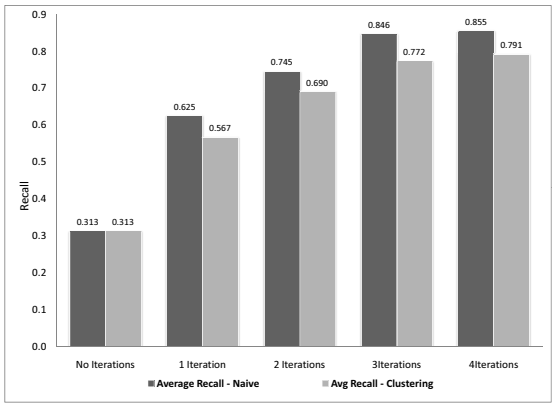
From each group an image is selected and matched against the remaining images of the same group. The decision about the visual similarity between two images is done based on the number of the interest points they share. It has been proposed that image matching approaches which are based on finding interest points correspondences provide a precision of 1 by using 5 common interest points as a threshold [7][10]. To exactly determine this threshold, we matched the input images to a group of 2000 images covering different categories. The results showed that a matching threshold of 4 common interest points produced zero false positives (precision of 1%) for almost all input images (the results are not show due to space limitations).

We adopted this matching threshold (4 interest points) and used it evaluate the matching recall under different number of iterations. In the iterative approach the matching images are used as input for further iterative applications of the same matching algorithm. The selection of the input images for each matching iteration is done in two ways: a naive one which considers all images found by a previous iteration, and a clustering based method which uses one representative image from each cluster as input. We used *agglomerative* hierarchal clustering to build the clusters. Agglomerative clustering is convenient to our case since we don't have to fix the number of the clusters and it is efficient since the number of the images that will be clustered before each matching iteration is small. After building the clusters, a random selection of a representative image from each cluster is applied to determine the inputs of the next matching iteration.

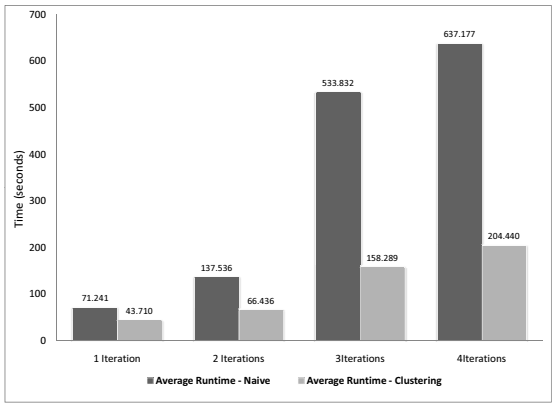
Fig. 6a shows the average matching recall for 0 (corresponds to a single matching without iterations) to 4 iterations. It can be seen, that the iterative approach results in significant recall improvement with an increasing number of iterations. Furthermore, the experiments show that for 76% of the cases 3 iterations were enough to achieve the maximum recall. On the other hand, the clustering based approach shows slight drop in the average recall compared to the naive approach, however, it provides better performance (Fig. 6b). The average runtime of the clustering approach is less than that of the naive approach and this difference increases by an increasing number of iterations. For example, by four iterations the clustering based approach is more than three times faster than the naive approach with a loss of about 0.065 in the recall.

4.2 Quality of Recommended Tags

The performance of the proposed annotation approach is evaluated by a leave-one-out cross-validation (LOOCV) approach: using a human-produced ground truth set of tagged images we compare the tags automatically generated with the actual ones. For this purpose, we created a dataset of 100 images and annotated them manually. Each image is geotagged and described with 21 tags on average. For each test image the annotation approach is applied and the pre-



(a) Average recall at different matching iterations



(b) Clustering-based vs. naive iterative matching: Average matching runtime needed by each approach at different matching iterations

Figure 6: Evaluation of iterative image matching approach

cision and recall are calculated by comparing the produced tags to the reference tags. After that, the average of precision and recall are calculated for the total set of test images. To show the effect of the produced ranking, the average precision and recall are demonstrated at different annotation lengths.

Word Importance / User Related Factor	
Weighted Voting/ -	TF-IDF/ -
Weighted Voting/$U_{\text{all users}}$	TF-IDF/$U_{\text{all users}}$
Weighted Voting/$U_{\text{users of w}}$	TF-IDF/$U_{\text{users of w}}$

Table 1: Configurations of the annotation approach

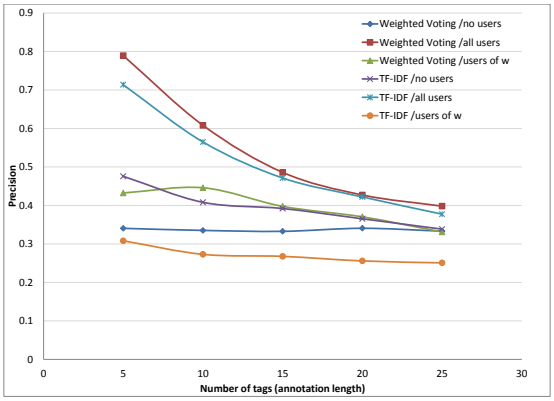


Figure 7: Average precision under different configurations and using ground truth of 100 images and with 0 iterations

The annotation approach can be configured according to two parameters. First, the word importance which can be calculated in two different ways: a weighted voting (Equation 3) and TF-IDF based approach (Equation 6). The second parameter is the user related factor U which can also can be calculated in two ways (Equations 7 and 8). In total, the two parameters can be combined in 6 different ways (Table 1).
Fig. 7 and 8 show that combination between the weighted

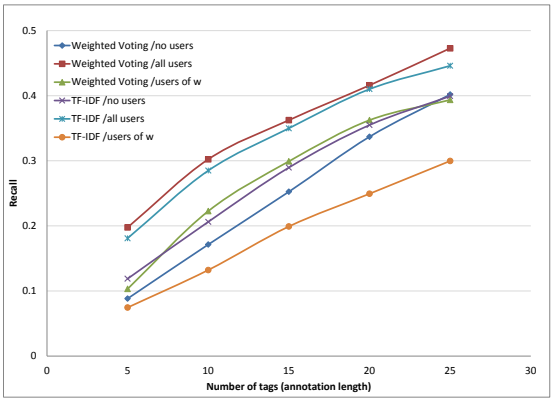


Figure 8: Average recall under different configurations and using ground truth of 100 images and with 0 iterations

voting and the user factor $U_{\text{all users}}$ provides the best recall and precision. This is a followed by TF-IDF combined with the same user factor. Additionally, we evaluated the annotation approach with a group of 88 images taken from NUS-WIDE dataset. NUS-WIDE database contains about 50 thousands geotagged images. Each image in the database is annotated with user tags in a uncontrolled manner, thus, the provided tag are somehow noisy. We cleaned a subset of the tags by removing irrelevant ones such numbers, stop words and user names. Fig. 9 and 10 show how the annotation performance using the weighted voting approach without user consideration under 0 and 1 iteration. The results show a clear improvement in the precision and recall when the iterative approach is applied. This is due to the fact, that the iterative matching approach results in finding additional user-tagged images. Consequently, the size of the candidate tags set as well as the accuracy of the produced tag ranking increase.

5. CONCLUSIONS

In this paper we presented an approach for automatic image annotation. It exploits geotagged images contributed

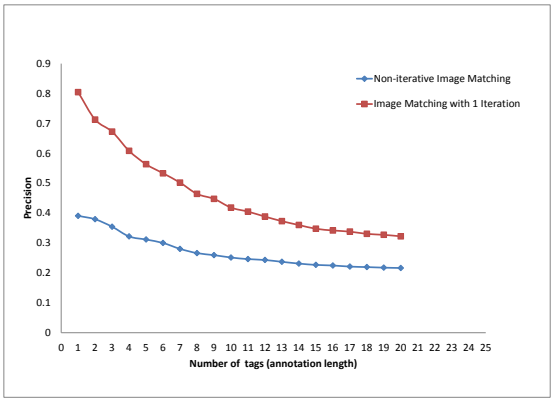


Figure 9: Average precision for a subset of NUS-WIDE dataset

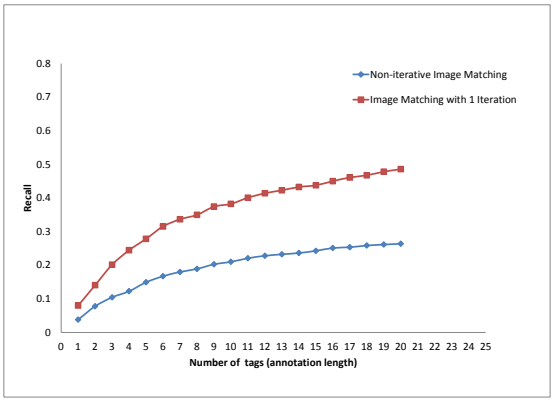


Figure 10: Average recall for a subset of NUS-WIDE dataset

to the web to automatically annotate new untagged images captured in the same location. We showed how user contributed images can be geographically filtered, visually matched, and how their tags can be used to annotate new images. The extracted information are combined in a flexible statistical model which is able to provide an image-word relevance score. A proposal is made to improve the visual matching by applying additional matching iterations and a solution for improving the performance based on image clustering is provided. Finally, the quality of the produced tags is evaluated by a human created ground truth. In our future work, we will consider performing further evaluation scenarios on larger datasets. Additionally, we aim to extend the probabilistic model to consider further image-word, image-image and word-word relations. We will also consider enhancing the performance of the provided approach so that it can fit an online system.

6. REFERENCES

[1] Y. Avrithis, G. Tolias, and Y. Kalantidis. Feature map hashing: Sub-linear indexing of appearance and global geometry. In *in Proceedings of ACM Multimedia (Full paper) (MM 2010)*, Firenze, Italy, October 2010.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110:346–359, June 2008.

[3] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*, Santorini, Greece., July 8-10, 2009.

[4] R. L. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. *IEEE Trans. on Knowl. and Data Eng.*, 19:370–383, March 2007.

[5] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40:5:1–5:60, May 2008.

[6] J. Eakins, M. Graham, and J. I. S. C. T. A. Programme. Content-based image retrieval, 1999.

[7] G. J. F. Jones, D. Byrne, M. Hughes, N. E. O'Connor, and A. Salway. Automated annotation of landmark images using community contributed datasets and web resources. In T. Declerck, M. Granitzer, M. Grzegorzek, M. Romanelli, S. M. Rüger, and M. Sintek, editors, *SAMT*, volume 6725 of *Lecture Notes in Computer Science*, pages 111–126. Springer, 2010.

[8] L. Juan and O. Gwun. A comparison of sift, pca-sift and surf. *International Journal of Image Processing (IJIP)*, 3(5), 2010.

[9] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:506–513, 2004.

[10] Y. Ke, R. Sukthankar, and L. Huston. Efficient near-duplicate detection and sub-image retrieval. In *ACM Multimedia*, volume 4, page 5, 2004.

[11] X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 11(7):1310–1322, 2009.

[12] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, and S. Ma. Dual cross-media relevance model for image annotation. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 605–614, New York, NY, USA, 2007. ACM.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, November 2004.

[14] M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *In Proceedings of the 15th International Conference on Multimedia (MM2007*, pages 631–640. ACM, 2007.

[15] G. Qiu. Image coding using a coloured pattern appearance model. In *Visual Communication and Image Processing*, 2001.

[16] T. Tachibanaya. Description of exif file format. *URL http://park2. wakwak. com/tsuruzoh/Computer/Digicams/exif-e. html. February*, 2001.

[17] L. Wu, L. Yang, N. Yu, and X.-S. Hua. Learning to tag. In *18th International World Wide Web Conference*, pages 361–361, April 2009.

[18] N. Zhou, W. K. Cheung, G. Qiu, and X. Xue. A hybrid probabilistic model for unified collaborative and content-based image tagging. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33:1281–1294, July 2011.