Neurocomputing 111 (2013) 144–153

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Tagging photos using users' vocabularies[☆]

Xueming Qian*, Xiaoxiao Liu, Chao Zheng, Youtian Du, Xingsong Hou

*School of Electronic and Information Engineering, Xi'an Jiaotong University, Xianning Road, Xi'an 710049, China*

ARTICLE INFO

*Article history:*
Received 1 September 2012
Received in revised form
27 October 2012
Accepted 10 December 2012
Communicated by M. Wang
Available online 8 February 2013

*Keywords:*
Social community
Social media
Image tagging
GPS
Batch tagging
User
Flickr

ABSTRACT

Online social image share websites such as Flickr and Panoramio allow users to manually annotate their images with their own words, which can be used to facilitating image retrieval and other image applications. The smart-phones have made it possible for users to capture images as well as get the geographical ordinates. It is easily recognized and accepted that visually similar images captured in the same place or in the same period of time may be also relevant in contents. In this paper we propose a personalized photo tagging approach by using users' own vocabularies. It can recommend users preferred tags for their newly uploaded photos based on the history information in their social communities by modeling users' tagging habit. The fundamental idea of our approach is that we try to recommend tags to users by accumulating votes from the candidate images. The candidate images are selected in term of three factors: visual features, geographical coordinates and image taken time. Thus, the candidate images include visually similar images, images captured in the same geographical coordinates or in the same period of time. Based on these three factors, we implement seven experiments. Experimental results on a Flickr image collection of nearly 2 million images of 5607 users demonstrate the effectiveness of our approach. The experimental comparison shows that the three factors have certain effectiveness in image tagging. The image tagging approach by fusing the image taken time, GPS information, and visual features achieve satisfactory performance . The impacts of history information and the batch tagging behavior to the image tagging performances are discussed.

## 1. Introduction

With the prevalence of social multimedia in the 21st century, digital images have become more and more accessible to the general public. Many users of online photo services (such as Flickr [1] and Panoramio [2]) are willing to share their image with family, friends, and the online community at large. When users share their images, they usually give their own vocabularies to describe the contents of their images, and this is the process of tagging. The prevalence of social multimedia tagging is significantly reshaping the way people generate, manage, and search multimedia resources. The tags provide descriptors of the images, and allow the user to organize and index images' contents.

With rapid development in technologies related to digital imaging, digital cameras also bring with camera metadata embedded in the digital image files. Camera metadata records information related to the image capture conditions and includes values such as tags, date/time stamps, subject distance and geographical coordinates. We have observed that the images in the same user's collection have a strong semantic relationship. Specifically, the images taken in the same time period or geographical coordinates are usually related to the same event. So how we can utilize the metadata of user's image collection to contribute to automatic image tagging for social media users is the key point we investigate.

Imagine that a Flickr or Panoramio user has been to one tourist attraction and taken some pictures by his or her mobile phone. To share with friends or families, the user uploaded some pictures to the Internet. The contribution of our approach is that we can recommend tags to their pictures using their own vocabularies in order to save users' time in annotating their photos. It can facilitate users to share their photos to their social communities.

Different from other tagging methods, the brilliant idea of our approach is that we take users' tagging habit into consideration to realize the personalized services. Every user has its own habit to tag images. Even for the same image, tags contributed by different users will be of great difference. For example, one may prefer "sea" to "ocean", but others may not. However, the existing tagging approaches do not care users' annotation habit and they generate the same tags for all the users. For example, a Flickr user's grandson is named "Jayden". He has uploaded a lot of photos about his lovely baby and tagged these photos with the word "Jayden". Existing tagging approaches may tag these photos

[☆] This work is supported by NSFC, Nos. 60903121 and 61173109, and Microsoft Research Asia.

* Corresponding author.
*E-mail address:* qianxm@mail.xjtu.edu.cn (X. Qian).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.12.021

with the word “baby” by analyzing image content [3–5]. However, if we recommend “baby” for the user, he may not be satisfied. Hence the aim of this paper is to recommend tags based on users' vocabularies. Our aim is recommending these photos with the word “Jayden” rather than “baby”.

The contributions of this paper can be described as follows: (1) we analyze users' tagging behavior by exploring their social communities; (2) we develop a personalized tagging approach by recommending tags for users using their own vocabularies; and (3) the influences of geographical coordinates, taken time and visual features of a user uploaded photo are fused into a unified tagging framework and their influences to final tagging performances are systematically analyzed.

The reminder of this paper is structured as follows. In Section 2, we review the related work on image tagging. Our approaches are illustrated in Section 3. The experimental setup and performance are shown in Section 4. In Section 5, the conclusions and future work are given.

## 2. Related work

Neighbor voting algorithm is proposed by Li et al. for image retrieval, which tried to get the relevance scores of user contributed tags by accumulating votes from similar images [3]. The tag recommendation based on collective knowledge is proposed [4]. The authors measured the similarity between tags by their co-occurrence information in the data collection, and used the top similar tags as recommendations. However, this kind of recommendation is based on single modality of tag co-occurrence on the whole dataset. Learning to tag formulated the recommendation as a learning to rank problem and combine three kinds of correlation (tag co-occurrence, tag visual correlation, and image conditioned tag correlation) to generate the ranking [5]. To enhance the descriptive ability of the existing tags and facilitate image retrieval, Yang et al. proposed a tagging approach, which aims at mining properties of tags such as shape, location, texture pattern, and color [6]. This approach can reduce the semantic gaps in tag based image retrieval. The relationship among tags can be modeled by a connected graph, the tagging can be converted to a graph based optimization problem [25,27]. A semi-automatic tagging scheme that can facilitate users in album tagging is proposed in [7]. The authors Liu et al. use a constrained affinity propagation algorithm to achieve the tradeoff between manual efforts and tag performance.

Various methods are intended to automatically annotate images. There is work on learning mappings from visual features to semantic labels in the machine learning communities and image processing [8,9]. The methods take a set of labeled images as input and learn which low level visual features correspond to higher level semantic labels. Then the mapping can be applied to suggest labels for unlabeled images based on visual features alone. Except for automatic image annotation, assistive tagging [10], namely tagging by combining human's intelligence and computer's computation power has drawn a lot of attention too. Wang et al. categorize existing assistive tagging into three paradigms: (1) tagging with data selection and organization; (2) tag recommendation; and (3) tag processing. For a more detailed account of content-based analysis in the field of image annotation we think of the ESP game [11]. ESP game is a tool for adding meaningful labels to images using a computer. Users suggest tags for photos that appear on their screen and earn points when suggesting the same tags as another player.

Image annotation or tagging research has also focused increasingly upon geo-tagging. Yahoo has released a product called Zonetag which offers geocoding (or geo-tagging) for the Flickr photos [9]. If camera phone is not GPS enabled, it will use cellular tower locations to approximate your coordinates or estimate image's location from its appearance [24] and tag your Flickr photos with that information [12,13]. Moxley et al. present a Spirit-tagger tool that mines tags from geographical and visual information [14]. These annotations are derived from image similarities constrained to a geographical radius, and a comparison of the local frequency in terms of their global frequency is used to weigh terms that occur frequently in a local area. In [15], a world-scale tag suggestion system is presented which employs a database of one million geo-tagged images in order to provide annotations for input photographs taken anywhere in the world. The first step involves prediction of geographical coordinates of the input image using the K-nearest-neighbor approach as in [16] that the user can choose to refine. Tag-cloud based suggestion systems are proposed recently by Joshi et al. [17,18], [17] is a preliminary tag-cloud suggestion system, while [18] constructs and evaluates the performance of multisource (public source, large-scale community source and personal source) location-driven tag-clouds as tag suggestion system [18]. In [21], both visual feature and geo-location of each image are fused to recommend image content related labels. The visual feature uses the probabilistic canonical correlation to predict the possible labels. The sensing location of an image from user's mobile terminal is mapped into the world-scale map grid to predict the candidate geo-tags. Tags are of great significance to image retrieval and image understanding. In [22], Wang puts forward a diverse relevance ranking image search scheme based on contents of images and their associated tags. This approach can avoid irrelevant or identical search results of existing tag-based ranking method. Ref. [23] finds a way to summarize web videos. The authors first localize the tags to video shots and then match shot-level tags with the query to identify key-shots. In [28], Li et al. presents a novel solution to the annotation of specific products in videos by mining information from the web. They use visual signatures to annotate video frames which are built based on the bag-of-visual-words representation of the training data. These data is collected by simultaneously leveraging Amazon and Google image search engine.

Qian and Hua model all the tags by a full connected graph [25]. They view tag enrichment as a combinational optimization problem. Graph cut based tag enrichment approach is proposed to determine the relevant tags. Tag enrichment is actually a graph cutting process. Each of the tags is either cut or kept with respect to the smooth term and data penalty term. Min-cut/Max-flow algorithm is resorted to find the optimal tag list for the input image. Moreover, in [26], Qian et al. carry out tag filtering by using similar compatible principles. This approach determines the ranks of user annotated tags by maximizing the compatible value of changing the labels of the tags from irrelevant to relevant at each step.

From above analysis we find that the existing image tagging approaches are on tagging image by using content relevant tags. However, to our knowledge it is the first time that we develop a personalized users' photo tagging approach using their own vocabularies based on user's history information. The history information includes three aspects: geographical coordinates, taken time, and visual features of user's image collection. The influences of the three aspects to the tagging performances are discussed.

## 3. Our approaches

### 3.1. Problem formulation

First, we introduce some notations. Let $I$, $T$, $P$, and $D$ denote the image collections, the set of tags, GPS locations and image taken dates of a user $u$ respectively. Let $M$ denote the number of the total uploaded images by the user $u$. We have the user's history information $H=\{I,T,P,D\}$ with

$$I=\{I_i\}_{i=1}^{M}, \quad T=\{T_i\}_{i=1}^{M}, \quad P=\{P_i\}_{i=1}^{M}=\left\{(x_i,y_i)\right\}_{i=1}^{M}, \quad D=\{D_i\}_{i=1}^{M} \tag{1}$$

where $\mathrm{I_i}$ means the ith image, $\mathrm{T_i}$ is the tags of $\mathrm{I_i}$, $\mathrm{P_i}$ is the GPS location of $\mathrm{I_i}$, and $D_i$ is the taken date of $\mathrm{I_i}$. $T_i=\emptyset$ means no tags are provided by the user for the ith image. $P_i=\emptyset$ means no GPS locations are assigned to the ith image.

The main information of the ith image $I_i$ can be a vector with six elements $s_i=\{u,p_i,d_i,\tau_i,z_i\}$:

(1) u is the user's name;
(2) $\mathrm{p_i}$ is the position the image $\mathrm{I_i}$ is taken;
(3) $\mathrm{d_i}$ is the taken date of the image I;
(4) $\tau_\mathrm{i}$ is the tag set we recommend to the image $\mathrm{I_i}$;
(5) $\mathrm{z_i}$ are the visual features of the image $\mathrm{I_i}$.

We call the user $u$'s image that we want to recommend tags to as input image. The input image is a newly uploaded image by user $u$, it has the GPS locations while has not been annotated by the user.

Among all these main information, user's name, taken position, taken time and visual features are the most useful ones to our methods. So the information of the input image can be written shortly as $s=\{u,p,d,\tau,z\}=\{u,(x,y),d,\tau,z\}$. Before utilizing our tagging method, the tag set $\tau$ of input image is empty. The proposed approach recommends tags for the newly uploaded image according to the users' history information $I$, $T$, $P$, and $D$.

### 3.2. Overview of our approaches

For tagging an input image $s=\{u,p,d,\tau,z\}=\{u,(x,y),d,\tau,z\}$, different from other tagging methods that highlight tag correlation between tags. We first use $p$, $d$, $z$ to search GPS neighbors, time neighbors and visual neighbors, respectively. Then we recommend tags for user $u$ by extracting user's own vocabularies in candidate image neighbors.

The detailed steps of our approach are as shown in Fig. 1. Firstly, we search GPS neighbors, time neighbors and visual neighbors from user's history information $H$ for the input image. The GPS neighbors are images sharing the same geographical ordinates with the input image. The time neighbors and input image are taken in the same period of time. The visual neighbors are visually similar images with the input image. Secondly, we collect all the tags of these candidate neighbors and count their appearing times by tags voting. These tags are user oriented, so we call them user's vocabularies. Finally, we annotate the image with the tags appear most frequently.

### 3.3. Searching GPS neighbors

In this paper, we recommend user related tags for the image according to users' history information of $H=\{I,T,P,D\}$. We choose the user's own images as GPS neighbors. Only in this way can we tag user's image with his or her own vocabularies. We determine whether the image $I_i$ is GPS neighbor of input image or not by comparing its GPS location $(x_i,y_i)$ with $(x,y)$:

$$G(i)=\begin{cases}1 & \text{if } sg(x_i-x,\alpha)=0 \quad \text{and} \quad sg(y_i-y,\alpha)=0 \\ 0 & \text{otherwise}\end{cases} \tag{2}$$

where $sg(x,\alpha)=0$, if the integer portion and the first $\alpha$ decimal places of $x$ are all 0; otherwise $sg(x,\alpha)=1$. $G(i)=0$ means that the image is not a GPS neighbor, while $G(i)=1$ means it is a GPS neighbor. In this paper, the total number of GPS neighbors in these $M$ images is defined as $N_G=\sum_{i=1}^{M}G(i)$. $N_G=0$ means that there is no GPS neighbors for the newly uploaded image taken at $(x,y)$. This is the case for the new users or the users capture photos at some new places. In these cases, the visual content or the taken time of the image $I_i$ can be utilized for tag recommendation. $N_G\neq 0$, which means that the user has already uploaded images taken in the same place with the newly uploaded image. $\alpha$ is a parameter that indicates the accuracy of geographical coordinates. When $\alpha$ is 5, the position of $(x,y)$ and $(x_i,y_i)$ are less than 1.1 m apart. We will discuss the influence of parameter $\alpha$ in Section 4.1.

### 3.4. Searching time neighbors

We determine whether the image $I_i$ is time neighbor of input image or not by comparing the taken time $d_i$ with the taken time $d$ of the input image

$$T(i)=\begin{cases}1 & \text{if } t(d_i,d)\le\beta \\ 0 & \text{otherwise}\end{cases} \tag{3}$$

where $t(d_i,d)$ means the time intervals between $d_i$ and $d$. $T(i)=0$ means that the $i$th image $I_i$ is not a time neighbor, while $T(i)=1$ means it is a time neighbor of input image. In this paper, the total number of time neighbors in these $M$ images is defined as $N_T=\sum_i^M T(i)$. $N_T=0$ means that there is no other images taken in the same date $d$ with the newly uploaded input image. We compare the tagging performances of $\beta$ as 1 month, a week and a day in our experiments in Section 4.1.

### 3.5. Searching visual neighbors

For each image in the user's image collection $I$, we compare its low level features with the input image. Features $z$ in $s=\{u,p,d,\tau,z\}=\{u,x,y,d,t,z\}$ of our approach are described as by the grid based color moment and the hierarchical wavelet packet descriptor.

Color feature has been proved to be the most GPS-informed feature [16,24]. It is used as global feature representation for the image in our method. An image is divided into four equal sized blocks and a centralized image with equal-size. For each block, a 9-D color moment is computed, and thus the dimension of color comment for each image is 45. The 9-D color moment of an image segment is utilized, which contains values of mean, standard deviation and skewness of each channel in HSV color space.

Texture feature has been shown to work well for texture description of image and for scene categorization and image recognition [19]. The texture feature in our method is described by hierarchical wavelet packet descriptor (HWVP) [20,29]. A 170-D HWVP descriptor is utilized by setting the decomposition level to be 3 and the wavelet packet basis to be DB2.

The visual similarity between images is measured by the Euclidean distance of two images as follows:

$$D(i)=\|z_i-z\| \quad i=1,2,\ldots,M \tag{4}$$

where $z_i$ and $z$ are the low-level feature of the image $I_i$ and the input image. In this paper, we rank the distances in ascending order and select the top ranked 10 images as its visual neighbors under the constraints that the visual similarity of two images are sufficient large.

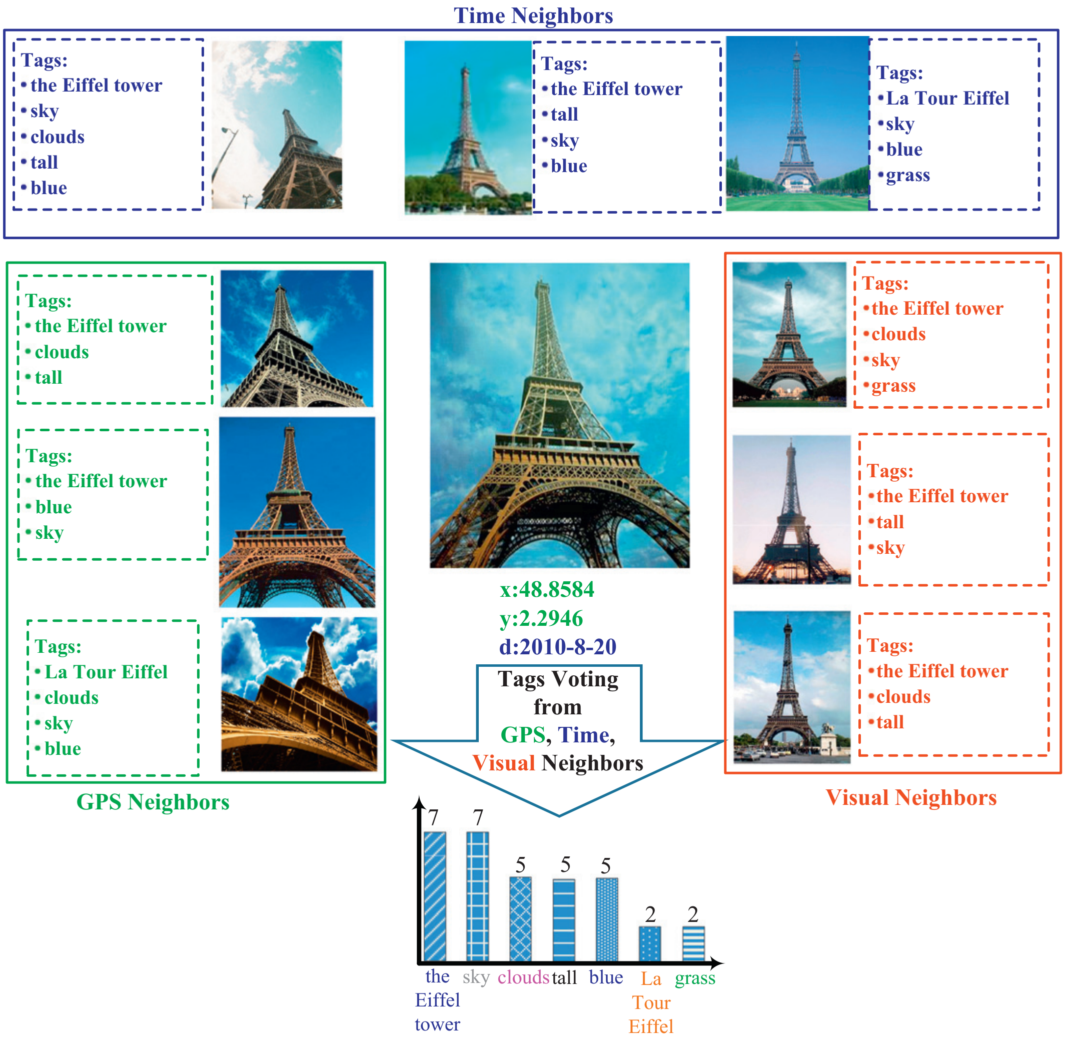


**Fig. 1.** Tagging photos using user's own vocabulary for the newly uploaded image by using the taken time, visual, and GPS information.

### 3.6. *Tags voting*

We use the tags appeared in the image neighbors to annotate the newly uploaded image by ranking repetition times of tags of the GPS, time and visual neighbors.

## 4. Experiments

### 4.1. *Experimental comparison*

We use three factors to help tagging images—geographical ordinates, taken time and visual features. In order to evaluate the three factors' influences on image tagging, we implement seven different experiments and compare their performances. These seven experiments are (1) UG, (2)UT, (3) UV, (4) UGV, (5)UGT, (6)UTV, and (7) GTV. The relationship between these seven approaches is illustrated in Fig. 2.

For tagging an input image $s=\{u,x,y,d,t,z\}$, UG only the geographical information $(x,y)$ to find its GPS neighbors. UT only utilizes the taken time $d$ to find time neighbors. UV only processes the image features $z$ of $s$ to find visual neighbors. UGV uses the geographical information $(x,y)$ and image features $z$ of $s$ to find GPS&Visual neighbors. UGT uses geographical information $(x,y)$ and the taken time $d$ to find GPS&Time neighbors. UTV makes use of taken time $d$ and visual features $z$ of $s$ to search for Time&Visual neighbors. GTV combines all the three factors—the geographical information$(x,y)$, taken time $d$ and visual features $z$ of $s$ together to find GPS&Time&Visual neighbors. In GTV, we use approach UGT to find GPS&Time neighbors. Then we search visually similar images among GPS&Time neighbors. And then propagate the top ranked tags of the neighbors to the input image respectively.

### 4.2. *Dataset*

In order to evaluate the performance of our methods, we randomly crawled more than 6 million images together with their tags from the image sharing site Flickr.com through its public API. The initial data includes 6,715,251 images uploaded by 7387 users and their related files recording the information of tags and geographical ordinates. We remove the information of images that have no tags and no geographical ordinates. We have made a statistic about the number and percentage of images that have tags, GPS or both. The result is shown in Table 1.

As we can see in Table 1, the remaining data with GPS and tags contains 1,903,089 images uploaded by 6581 users. That is to say, most users have the habit to give their images tags or geographical ordinates. For every user, we choose the image uploaded most recently as input image for testing, and we view other images as the training set of this user. Among these testing images, about 14.8% of them are in batch tagging mode. We remove these batch tagging images to avoid their influence on our tagging methods. So it turns

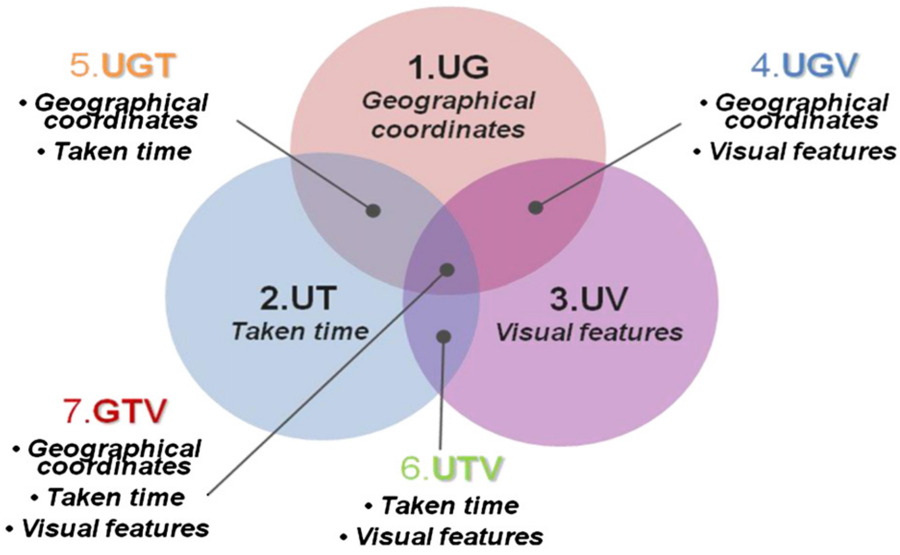


**Fig. 2.** The relationship of the seven tagging approach for user newly uploaded photos: UG, UT, UV, UGV, UGT, UTV, and GTV by using geographical coordinates, taken time, and visual features.

**Table 1**
Flickr users and their uploaded images in our experiment.

| Total | With tag | With GPS | With GPS+tag | |
|---|---|---|---|---|
| *User* | | | | |
| Number | 7387 | 7276 | 7276 | 6581 |
| Percentage | 100 | 98.50 | 98.50 | 89.09 |
| *Image* | | | | |
| Number | 6,715251 | 5,317,909 | 2,144,661 | 1,903,089 |
| Percentage | 100 | 79.19 | 31.94 | 28.34 |

out that there are 5607 images for testing the performances of the proposed tagging approaches.

### 4.3. Criteria of performance evaluation

For the input image, the user has annotated $o$ tags $t=\{t_1,t_2,...,t_o\}$. When the input image has been uploaded by this user, $o$ is an invariant value. For each input image we recommend $r$ tags $\tau=\{\tau_1,\tau_2,...,\tau_r\}$ to the user. By comparing tags in these two sets $t=\{t_1,t_2,...,t_o\}$ and $\tau=\{\tau_1,\tau_2,...,\tau_r\}$, we find that some are the same with the original tags but the others are not. In this paper we use Recall, Precision and F1 to measure tagging performance of a test image, which are defined as follows:

$$Recall=\frac{c}{c+m}100\%=\frac{c}{o}100\% \tag{5}$$

$$Precision=\frac{c}{c+f}100\%=\frac{c}{r}100\% \tag{6}$$

$$F1=\frac{2\times Recall\times Precision}{Recall+Precision}100\% \tag{7}$$

where $c, f$, and $m$ are the number of correct, false and missed tags. We use the average recall (AR), average precision (AP) and average F1 (AF) of 6581 users under different $r$ for evaluating tagging performance.

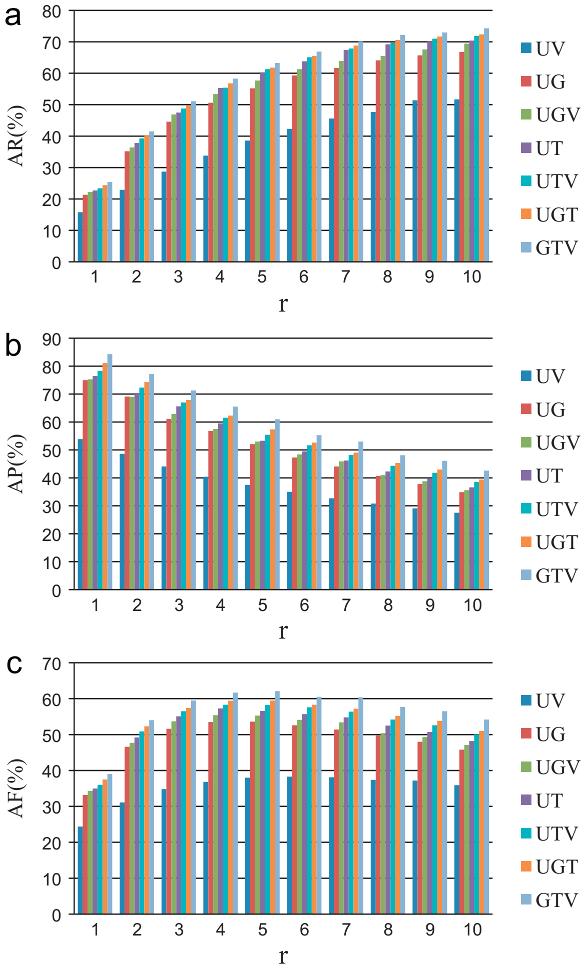
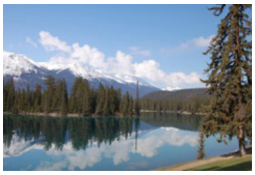


**Fig. 3.** The (a) AR, (b) AP and (c) AF values of 5607 users for the seven approaches GTV, UGV, UGT, UTV, UG, UT, and UV when the recommended tag number $r$ is in the range of [1,10]. The parameters are $\alpha=5$ and $\beta=1$ month (a) AR under different tag number $r$, (b) AP under different tag number $r$ and (c) AF under different tag number $r$.

**Table 2**
Exemplar images for showing the performances of these seven approaches. This table includes the photo, initial tags the user gave (INIT) and the tags recommended by these seven methods (GTV, UGT, UTV, UT, UGV, UG and UV).

| Photos | Recommended tags | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | INIT | GTV | UGT | UTV | UT | UGV | UG | UV |
| | old sky germany monastery romania biserica sibiu cladiri cer hermannstadt | sky old germany monastery romania biserica sibiu | sky old germany monastery romania biserica sibiu building | sky germany monastery romania cer fall autumn | sky old germany monastery romania biserica cer fall autumn | germany monastery romania sibiu cladiri biserici | germany monastery romania sibiu cladiri maramure biserici building | old sky blue building maramure biserici |
| | flower nature natura white green comanesti | flower nature natura white green | flower nature white natura green sunrise | flower flowers nature natura | flower flowers nature natura sunrise morning | flower grass nature natura mountain | flower grass nature natura white green mountain sunrise | flower nature natura bee insecte floare |
| | net boat shrimp line anchor beached d200 fishingboat dock beaufort overhaul chasitybrooke beuafortnc | boat shrimp anchor d200 beached dock beaufort net overhaul sky | boat shrimp anchor d200 beached dock beaufort net overhaul sky | boat boats shrimp beaufort d200 sky sun dock | boat boats shrimp beaufort d200 sky sun water clouds dock | boat fishing boat dock sunfish d200 beaufort | sky shrimp water clouds boat fishingboat dock sunfish d200 beaufort | sky boat blue clouds dock fishing |
| | winter sun snow mountains ice nature norway landscape 78°n nikon north glacier valbard dogsledding spitsbergen d40 | winter snow sun norway mountains nikon glacier | winter snow sun norway mountains nikon glacier dogsledding longyear | winter ice norway nikon north glacier isbre | winter ice norway nikon north glacier longyear isbre vonpost | winter snow mountains norway svalbard nikon longyear vonpost 78°n | winter snow mountains norway svalbard Spitsbergen longyear vonpost coal 78°n | winter snow ice nikon mountains coal mine traffic svalbard |
| | sunset sun nature colors norway landscape norge colors north arctic midnightsun nordland steigen leines | sunset nature norway landscape colors clouds | sunset nature norway landscape colors nordland clouds red | sunset nature norway landscape colors nordland clouds morning august | sunset nature norway landscape colors nordland clouds morning august water | sunset sunrise nature norway landscape colors nordland clouds water | sunset sunrise nature norway landscape colors nordland clouds red water | sunset sunrise landscape red clouds water |
| | pez candy sweet tizzy sweetcandy tz1 | candy sweet tizzy tz1 sugar | candy candies sweet tizzy tz1 sugar | candy sweet tizzy tz1 sugar light stars | candy sweet tizzy tz1 sugar light | candy sweet tizzy tz1 sugar light stars candies | candy sweet light stars lantern tizzy tz1 | candy sweet light stars lantern japanese |
| | flowers plants netherlands garden flora friesland buitenpost kruidhof 1530sigma | plants netherlands garden friesland buitenpost kruidhof | plants netherlands garden flora friesland buitenpost kruidhof | plants netherlands garden friesland building spring | plants netherlands garden friesland building spring bloom | plants netherlands garden friesland building | plants netherlands garden flora friesland green building buitenpost kruidhof | green grass plants building rain |
| | park family atlanta me bench dad | park atlanta me girl dad bench | park atlanta me girl dad hair sun bench | park atlanta girl dad sunshine bench emory | park atlanta dad sunshine bench girl fence hair emory | park atlanta dad fence emory | park atlanta dad sun jeans fence emory | park atlanta girl hair dad sun sunshine |
|  | lake canada reflection landscape nikon jasper alberta d40 | landscape nikon alberta canada d40 lake | landscape canadanikon alberta sky central d40 lake | sky canada landscape nikon farmland alberta d40 | lake sky alberta canada june 2008 landscape farmland d40 | canada central alberta landscape farmland nikon d40 | canada central alberta landscape nikon d40 farmland back country | sky clouds landscape trees farmland nikon d40 |

### 4.4. Results of our experiments

Fig. 3(a)–(c) shows respectively the AR, AP and AF values of the seven tagging approaches: UV, UG, UGV, UT, UTV, UGV and GTV of the 5607 users when the recommended tag number $r$ is in the range of 1–10. From Fig. 3, we find that UV is with lowest performances. It means that from the visual information the users' tag predication performances are not satisfactory. With the help of GPS information of the input image, better performances are achieved. Combining both the visual information over the geo-tag information, some improvements are made. The user' uploading time is also very useful for recommending correct tags. By combining the geo-tags, time and visual information, better tagging performances are achieved.

Fig. 3(a)–(c) shows the corresponding AR, AP and AF for the seven approaches. We can see that as the recommended tag number $r$ grows, the AP value drops, while the AR value increases. As $r$ grows, the correct recommended tag number $c$ may probably increases, because we have more opportunity to recommend right tags. For an input image, $o$ is an invariant value, $r$ is the denominator of Precision, so when $r$ grows, AP will drops.

In GTV, AR reaches up to about 26% when $r=1$, and at the same time the AP and AF are about 85% and 40% respectively. AR reaches up to about 75% when $r$ is 10, and at the same time the AP and AF are about 44% and 55%. The performance gaps between GTV and UTV, UT, UGV and UG are about one to two percentages. From Fig. 3 we find that the performances of using only the visual information (i.e. UV) of the photos are not very good in recommending user preferred tags. When $r$ is 5, Recall and Precision values of UV are all less than 40%. The AF values of UT are less than 40% with $r$ in the range [1,10]. While under $r=5$, AR of GTV reaches up to about 64%, and at the same time the AP and AF are about 62% and 63% respectively. That is to say, more than 50% recommended tags are matched with user annotated tags. It shows the effectiveness of the proposed tagging approaches.

Furthermore, we give the exemplar images and the recommended tags of the seven approaches: GTV, UGT, UTV, UT, UGV, UG and UV using correspondingly the photo's taken time, visual information and GPS information to show their effectiveness in Table 2. We recommend ten tags for the testing image with the initial tags the user gave (denoted INIT). Sometimes the recommended tag number is less than ten, under the case their neighbors are less than ten. The tags in red are the same with the user contributed tags. The blue ones are relevant with the image while the black ones are of no relationship with the input image. The specific tag number depends on the result of tags voting in candidate neighbors. By comparing the recommended tags of the seven approaches with the user generated tags, we find that GTV is with best performances while UV is with lowest performances.

### 4.5. Discussion

In the social media websites, different users have different history information. Some users are likely batch tagging their photos. Thus, in this section we give detailed discussions on the influences of GPS accuracy, time interval, history information and the batch tagging behavior to the users' photo tagging performances.

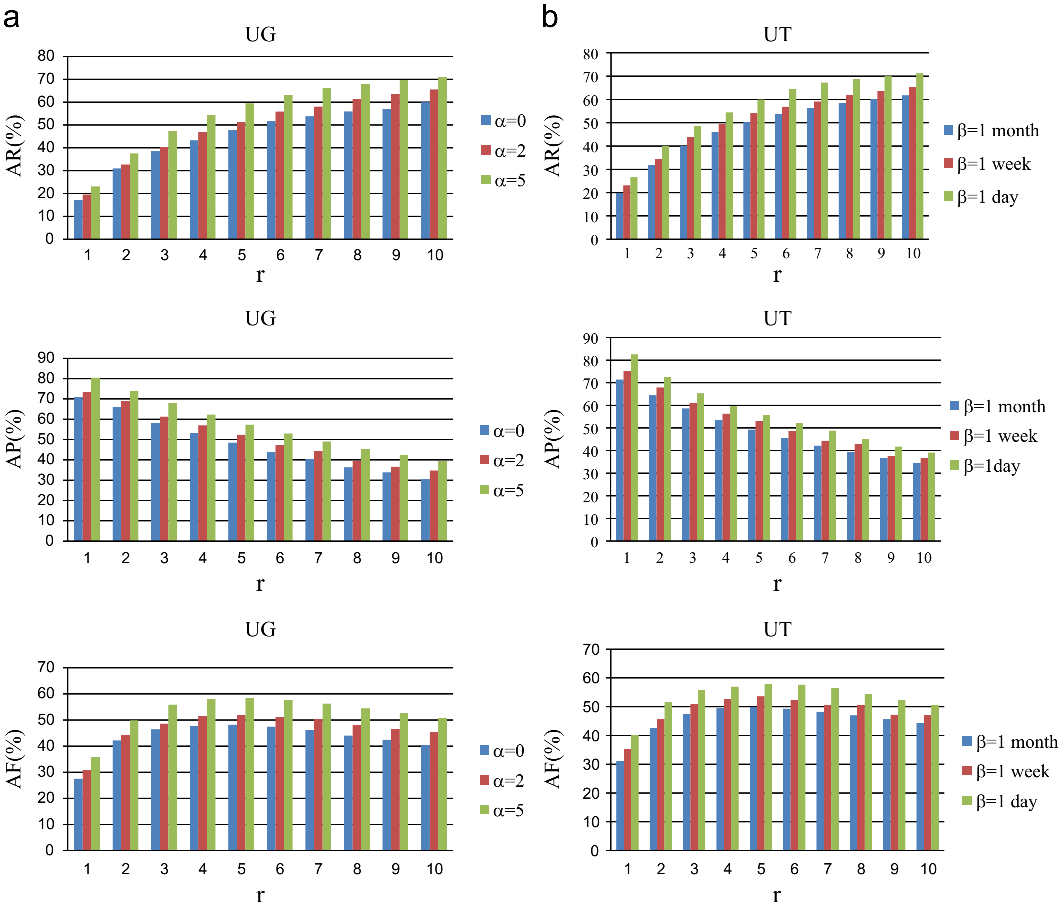


**Fig. 4.** The AR, AP and AF values of UG and UT under different tag number $r$: (a) UG under $\alpha=\{0,2,5\}$ and (b) UT under $\beta=\{1\text{ month}, 1\text{ week}, 1\text{ day}\}$.

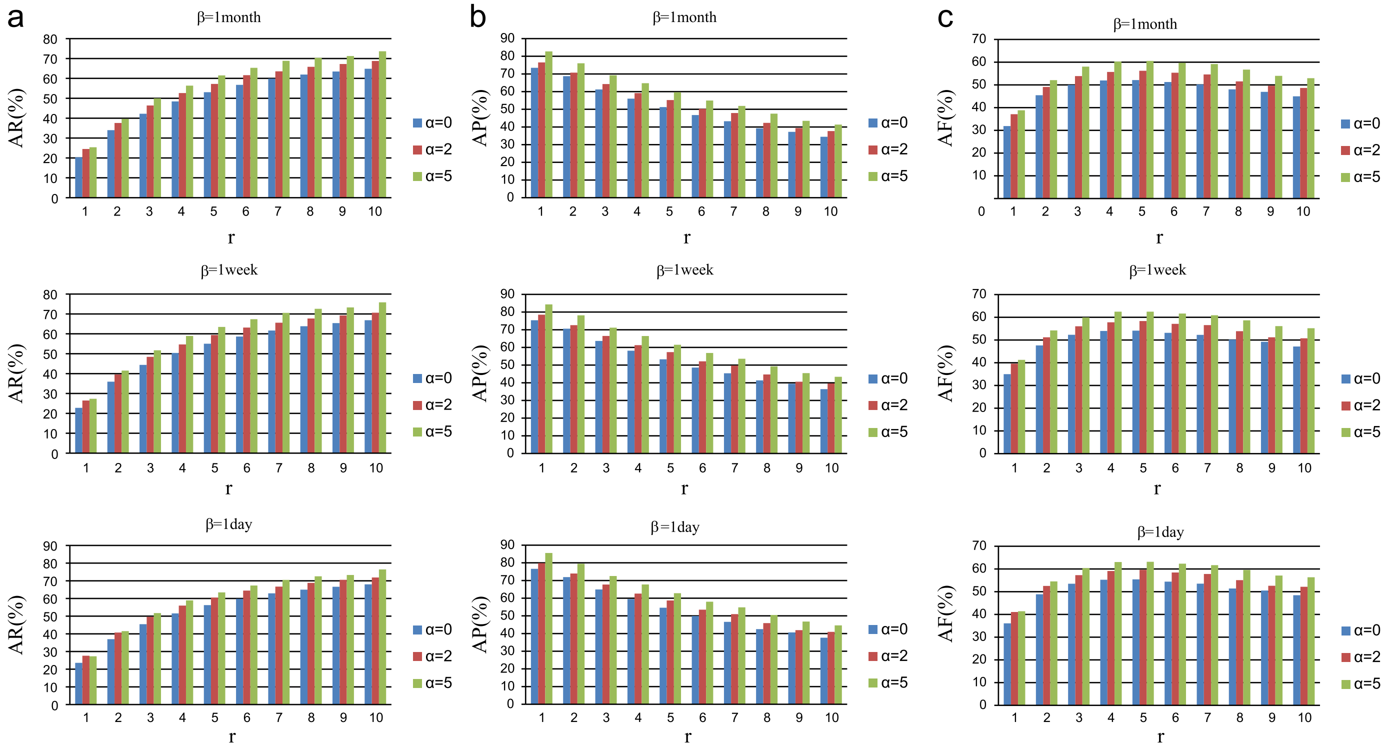


**Fig. 5.** The tagging performances of GTV under $\alpha=\{0,2,5\}$ with $\beta=\{1\text{ month, 1 week, 1 day}\}$ (a) AR values of GTV under $\alpha=\{0,2,5\}$ with $\beta=\{1\text{ month, 1 week, 1 day}\}$, (b) AP values of GTV under$\alpha=\{0,2,5\}$ with $\beta=\{1\text{ month, 1 week, 1 day}\}$ and (c) AF values of GTV under$\alpha=\{0,2,5\}$ with $\beta=\{1\text{ month, 1 week, 1 day}\}$.

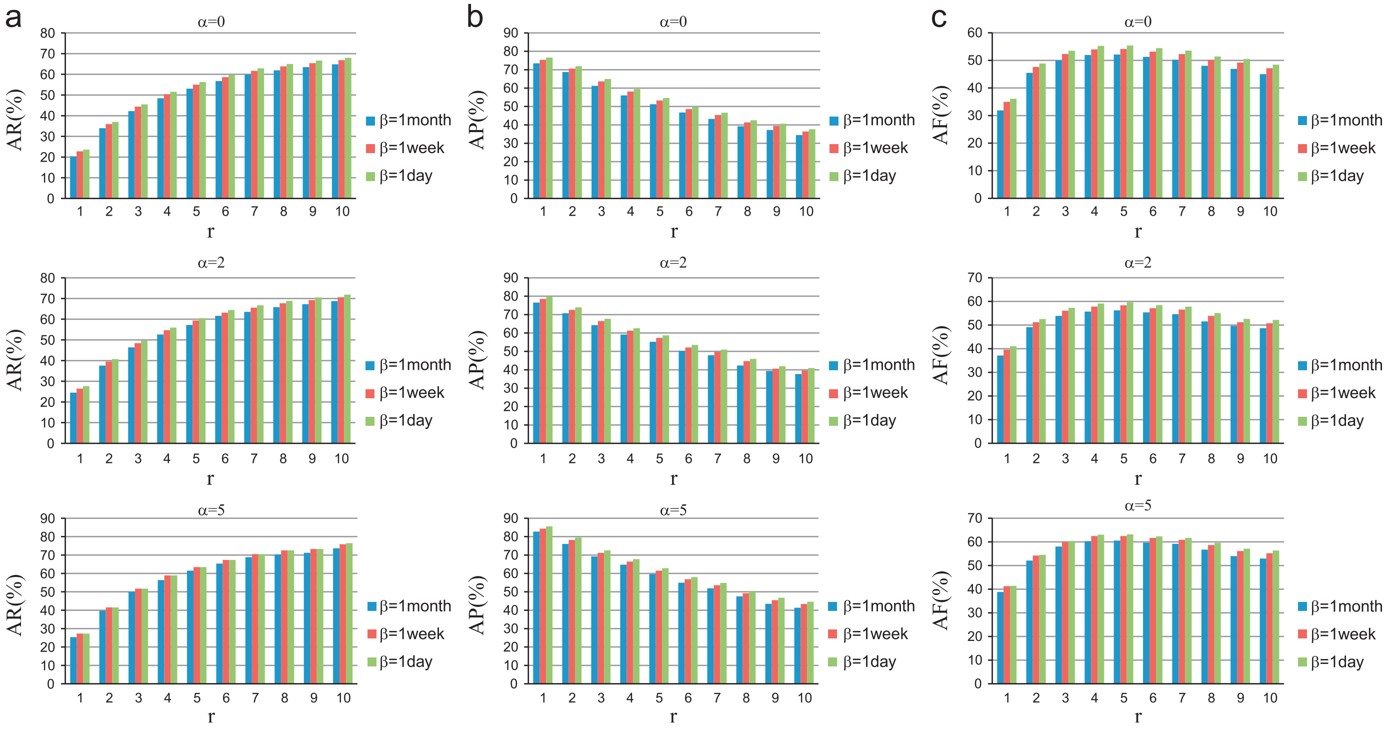


**Fig. 6.** The tagging performances of GTV under $\beta=\{1\text{ month, 1 week, 1 day}\}$ with $\alpha=\{0,2,5\}$: (a) AR values of GTV under $\beta=\{1\text{ month, 1 week, 1 day}\}$ with $\alpha=\{0,2,5\}$, (b) AR values of GTV under $\beta=\{1\text{ month, 1 week, 1 day}\}$ with $\alpha=\{0,2,5\}$ and (c) AF values of GTV under $\beta=\{1\text{ month, 1 week, 1 day}\}$ with $\alpha=\{0,2,5\}$.

### *4.5.1. The influence of parameters α and β*

In this part, we will discuss the influence of the parameters $\alpha$ and $\beta$ to the performances of proposed tagging approach. In order to show the influences of each aspect to the final tagging performance, we give the performances of GTV with fixed time interval and with various GPS accuracies (i.e. UG) and the GTV with fixed GPS accuracy (i.e. UT) and with various time intervals are provided in Fig. 4. In UG, $\alpha$ indicates the accuracy of geographical coordinates. In UT, $\beta$ is the time interval between the image $I_i$ and the input image. The impacts of $\alpha$ (under 0, 2 and 5) and $\beta$ (under 1 month, a week and a day) to tagging performances are shown in Fig. 4(a) and (b) respectively. As can

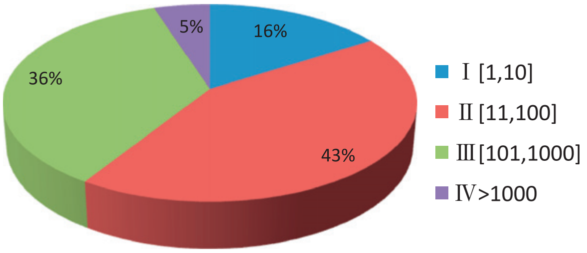


**Fig. 7.** The number of history images of the 5607 users. As it is shown, Class I–Class IV represent the users has no more than 10 images, 11–100 images, 101–1000 images, and more than 1000 images, respectively.

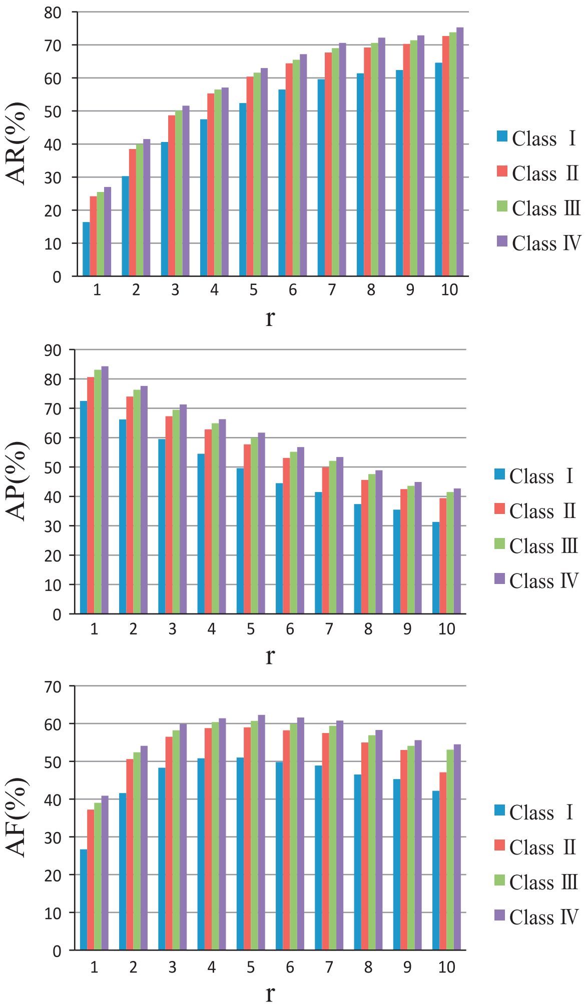


**Fig. 8.** The tagging performances of GTV under different number of history images. The users are divided into four categories based on the history image number.

be seen, the more accurate the geographical coordinates are or the shorter the time interval is, the better the AR, AP and AF are. Moreover, the comprehensive discussions for GTV under different $\alpha$ and $\beta$ values are shown in Figs. 5 and 6 respectively. As the experiments illustrated, GTV with short time intervals and more accurate geographical coordinates achieves better performance.

### 4.5.2. The influence of history image number on GTV

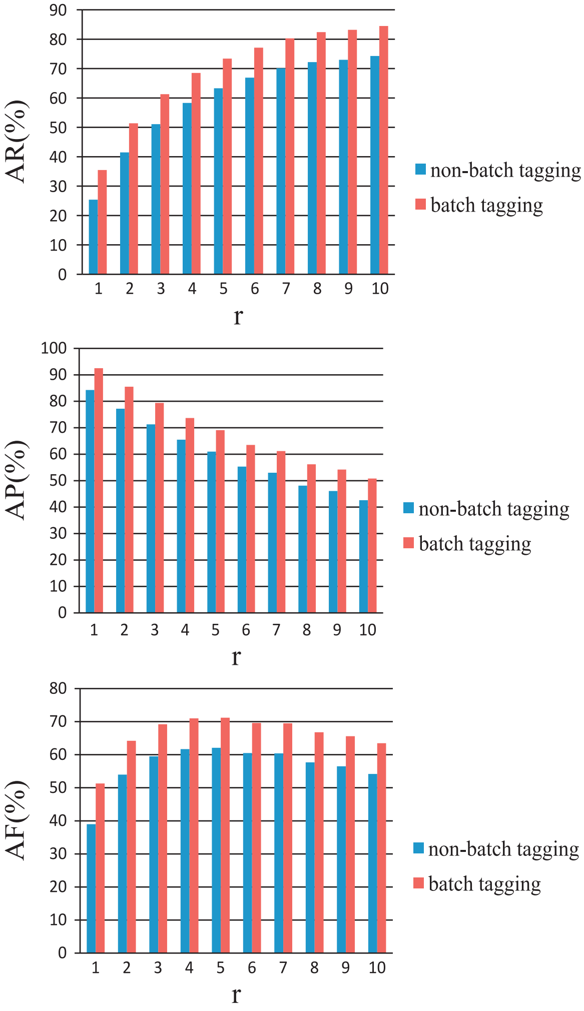


**Fig. 9.** The impact of batch tagging and non-batch tagging to image tagging performances of GTV.

It is obvious that it would be easier to find candidate neighbors with more history images. So we also set up experiments to analyze the relationship between the number of history images and the tagging performance. First we make a statistic about the image number in a user's history data. Then we categorize the 5607 users into four paradigms based on their history image number. The four classes are for the users with their uploaded image number in the range [1,10] (denoted Class I), [11,100] (denoted Class II), [101,1000] (denoted Class III), and $>1000$ (denoted Class IV).

Fig. 7 shows the percentage of users in each class. We can see that the percentages of the users of the four classes are about 5%, 43%, 36% and 16%. The results of GTV on different user class are illustrated in Fig. 8. The parameters of GTV are $\alpha = 5$ and $\beta = 1$ month. As it is shown in Fig. 8, when a user has no more than ten images, the performance of tagging method GTV lags behind users that have more images. With adequate images, our tagging results will be more precise.

### 4.5.3. The influence of batch tagging on GTV

In Fig. 9 the impacts of batch tagging (denoted batch tagging) or non-batch tagging (denoted non-batch tagging) behavior image tagging performances are discussed. It is rational that batch tagging images have higher performance. The reason is that it is easier to find GPS or time neighbors in batch tagged images. And these neighbors have exactly the same tags with the input image.

## 5. Conclusions and future work

In this paper, we develop a newfangled approach for image tagging. We make a good use of the geographical ordinates, time taken and the visual features of user shared images in their social communities. We propose a tagging approach for the newly uploaded photos using user's own vocabularies. Relevant annotations are highly dependent on geographical coordinates and time taken. What is more, experiments indicate that with more history images, our tagging results will be more precise. However, there is still much work to be done. First, we will dig deep on how to recommend tags to the user when there are no candidate image neighbors of the input image. Second, we will work on how to develop image content understanding with the help of geo-tagging.

## References

[1] Flickr, 〈http://www.flickr.com/〉.

[2] Panoramio, 〈http://www.panoramio.com/〉.

[3] X. Li, C.G. Snoek, M. Worring, Learning tag relevance by neighbor voting for social image retrieval, in: Proceedings of the MIR, 2008.

[4] B. Sigurbjörnsson, R.V. Zwol, Flickr tag recommendation based on collective knowledge, in: Proceedings of the WWW, 2008.

[5] L. Wu, L. Yang, N. Yu, X. Hua, Learning to tag, in: Proceedings of the WWW, 2009.

[6] Xian-Sheng Kuiyuan Yang, Meng Hua, Hong-Jiang Zhang. Wang, Tag tagging: towards more descriptive keywords of image content, IEEE Trans. Multimedia 13 (4) (2011) 662–673.

[7] Meng Dong Liu, Xian-Sheng Wang, Hong-Jiang Zhang. Hua, Semi-automatic tagging of photo albums via exemplar selection and tag inference, IEEE Trans. Multimedia 13 (1) (2011) 82–91.

[8] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, M. Jordan, Matching words and pictures, J. Mach. Learn. Res. 3 (2003) 1107–1135.

[9] J. Li, J.Z. Wang, Real-time computerized annotation of pictures, ACM Multimedia (2006) 911–920.

[10] Meng Wang, Bingbing Ni, Xian-Sheng. Hua, Tat-Seng Chua, Assistive tagging: a survey of multimedia tagging with human–computer joint exploration,, ACM Comput. Surv. 44 (4) (2012).

[11] Luis von Ahn, and Laura Dabbish, "Labeling Images with a Computer Game", In Proc. CHI, 2004, PP. 319–326.

[12] Zonetag. 〈http://zonetag.research.yahoo.com/〉.

[13] S. Ahern, M. Davis, D. Eckles, S. King M. Naaman, R. Nair, M. Spasojevic, J. Yang Zonetag: designing context aware mobile media capture to increase participation, in: Proceedings of Workshop on Pervasive Image Capture and Sharing, 2006.

[14] E. Moxley, J. Kleban,B.S. Manjunath: SpiritTagger: a geo-aware tag suggestion tool mined from Flickr, in: Proceedings of ACM Multimedia Information Retrieval (MIR), 2008.

[15] J. Kleban, E. Moxley, J. Xu, B.S. Manjunath, Global annotation on georeferenced photographs, in: Proceedings of the CIVR, 2009.

[16] J. Hays, A. Efros IM2GPS: estimating geographic information from a single image, in: Proceedings of the CVPR, 2008.

[17] D. Joshi, J. Luo, J. Yu, P. Lei, A. Gallagher, Rich location-driven tag cloud suggestions based on public, community, and personal sources, in: Proceedings of ACM International Workshop on Connected Media Mining, 2010.

[18] Dhiraj Joshi, Jiebo Luo, Jie Yu, Phoury Lei, Andrew C. Gallagher: Using Geotags to Derive Rich Tag-Clouds for Image Annotation. Social Media Modeling and Computing 2011: 239-256. http://dx.doi.org/10.1007/978-0-85729-436-4_11, springer-verlag London Limited 2011.

[19] B. Manjunath, W. Ma, Texture features for browsing and retrieval of image data, IEEE Trans. Pattern Anal. Mach. Intell. 18 (8) (1996) 837–842.

[20] X. Qian, G. Liu, D. Guo, Z. Li, Z. Wang, H. Wang, Object categorization using hierarchical wavelet packet texture descriptors, in: Proceedings of the ISM, 2009, pp. 44–51.

[21] A. Shimada, H. Nagahara, R. Taniguchi, Geolocation based image annotation, in: Proceedings of the ACPR, 2011, pp. 657–661.

[22] Kuiyuan Meng Wang, Xian-Sheng Yang, Hong-Jiang Zhang. Hua, Towards a relevant and diverse search of social images, IEEE Trans. Multimedia 12 (8) (2010) 829–842.

[23] Richang Meng Wang, Guangda Hong, Zheng-Jun Li, Shuicheng Zha, Tat-Seng Chua. Yan, Event driven web video summarization by tag localization and key-shot identification, IEEE Trans. Multimedia 14 (4) (2012) 975–985.

[24] J. Li, X. Qian, Yuan Yan Tang, L. Yang, and C. Liu, GPS estimation from users' photos, in: Proceedings of the MMM, 2013.

[25] X. Qian, X. Hua, Graph-cut based tag enrichment, in: Proceedings of the SIGIR, 2011, pp. 1111–1112.

[26] X. Qian, X. Hua, X. Hou, Tag filtering based on similar compatible principle, in: Proceedings of the ICIP, 2012, pp. 2349–2352.

[27] E. Moxley, T. Mei, B. Manjunath, Video annotation through search and graph reinforcement mining,, IEEE Trans. Multimedia 12 (3) (2010) 184–193.

[28] Meng Guangda Li, Zheng Wang, Richang Hong Lu, Chua. Tat-Seng, In-video product annotation with web information mining, ACM Trans. Multimedia Comput. Commun. Appl. 8 (4) (2012) 55:1–55:19.

[29] X. Qian, H. Wang, G. Liu, X. Hou, "HWVP: Hierarchical Wavelet Packet Texture Descriptors and Their Applications in Scene Categorization and Semantic Concept Retrieval", Multimedia Tools and Applications, May (2012), http://dx.doi.org/10.1007/s11042-012-1151-8.

**Xueming Qian** (M'10) received the B.S. and M.S. degrees in Xi'an University of Technology, Xi'an, China, in 1999 and 2004, respectively, and the Ph.D. degree in the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China, in 2008. He was awarded Microsoft fellowship in 2006. From 1999 to 2001, he was an Assistant Engineer at Shannxi Daily. From 2008 till now, he is a faculty member of the School of Electronics and Information Engineering, Xi'an Jiaotong University. Now he is an associate professor of the School of Electronics and Information Engineering, Xi'an Jiaotong University. He is the director of SMILES LAB. He was a visit scholar at Microsoft research Asia from Aug. 2010 to March 2011. His research interests include video/image analysis, indexing, and retrieval.

**Xiaoxiao Liu** received the B.E. degree from the Xi'an University of Post and Telecomunications, Xi'an, China, in 2011. She is currently sophpmore postgraduate with the School of Electronics and Information Engineering in Xi'an Jiaotong University, Xi'an, China. Now she is a MSD student at SMILES LAB.

**Chao Zheng** received the B.S. degrees in Xi'an Jiaotong University, Xi'an, China,in 2012. He attended XJTU's Information-Technology Talent Program (ITP), while he was in the undergraduate period. He was a visiting student at SMILES LAB from May 2010 to July 2012.

**Youtian Du** received B.S degree in department of electric engineering from Xi'an JiaoTong University, China in 2002, the Ph.D degree in department of automation from Tsinghua University, China, in 2008. He is currently an assistant professor of Xi'an Jiaotong University. His research interests include online social network, web image and video understanding, and machine learning.

**Xingsong Hou** received the B.S. degree in electronic engineering from North China Institute of Technology,-Taiyuan,China, in 1995, and the M.S. degree and Ph.D degree in the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China in 2000 and 2005, respectively. From 1995 to 1997, he was an Engineer with the Xi'an Electronic Engineering Institute in the field of radar signal processing. Now he is an associate professor of the School of Electronics and Information Engineering, Xi'an Jiaotong University. His research interests include video/image coding, wavelet analysis, sparse representation, sparse representation and compressive sensing, and radar signal processing.