Future Internet **2010**, *2*, 341-362; doi:10.3390/fi2030341

OPEN ACCESS

future internet

ISSN 1999-5903

www.mdpi.com/journal/futureinternet

Article

Can Global Visual Features Improve Tag Recommendation for Image Annotation?

Mathias Lux [1,*], Arthur Pitman [2] and Oge Marques [3]

[1] Institute for Information Technology, Klagenfurt University, Universitaetsstr, 65-67, 9020 Klagenfurt, Austria

[2] Institute for Applied Informatics, Klagenfurt University, Universitaetsstr, 65-67, 9020 Klagenfurt, Austria; E-Mail: arthur.pitman@uni-klu.ac.at

[3] Department of Computer & Electrical Engineering and Computer Science, Florida Atlantic University, 777 Glades Rd, Boca Raton, FL 33431, USA; E-Mail: omarques@fau.edu

* Author to whom correspondence should be addressed; E-Mail: mlux@itec.uni-klu.ac.at.

Received: 5 August 2010; in revised form: 21 August 2010 / Accepted: 26 August 2010 / Published: 27 August 2010

Abstract: Recent advances in the fields of digital photography, networking and computing, have made it easier than ever for users to store and share photographs. However without sufficient metadata, e.g., in the form of *tags*, photos are difficult to find and organize. In this paper, we describe a system that recommends tags for image annotation. We postulate that the use of low-level global visual features can improve the quality of the tag recommendation process when compared to a baseline statistical method based on tag co-occurrence. We present results from experiments conducted using photos and metadata sourced from the *Flickr* photo website that suggest that the use of visual features improves the mean average precision (MAP) of the system and increases the system's ability to suggest different tags, therefore justifying the associated increase in complexity.

Keywords: image retrieval; multimedia; metadata; folksonomies; tagging; image annotation; tag recommendation; visual information retrieval

1. Introduction

It has never been so easy and inexpensive to take pictures and subsequently store, share and publish them. Thanks to many recent advances in image compression, computer networks and web-based

technologies, accompanied by a significant reduction in hardware costs, our picture-taking habits have changed dramatically. We live in a world where more people have cameras than ever before; the average number of pictures taken per person is at least an order of magnitude higher than it used to be during the pre-digital era, and many of these pictures are uploaded to social sharing sites at an astounding rate to be viewed by an audience of millions. For example, Flickr (www.flickr.com), a well known web platform for storing, organizing and sharing photos, has more than 26 million members [1] and grows by more than 6000 photos uploaded each minute [2].

Despite all these advances in creating and storing images, the tasks of finding images of interest and retrieving them remain as challenging as ever. One of the main difficulties in image retrieval is the fact that most successful search engines are text-based and therefore rely on the presence of text (e.g., keywords) associated with the images to be able to properly retrieve them. In the case of social sharing sites, such keywords usually appear as *tags* associated with the images. In a perfect world, all images would have a reasonable number of user-generated tags, which would then enable other users to find and retrieve them. Unfortunately, in reality, only a fraction of the uploaded pictures are tagged with useful tags by their users, leaving an enormous number of (potentially good and interesting) pictures buried in a place that keyword-based search engines cannot reach.

Since the early 1990s, the desire to be able to locate and retrieve images regardless of textual metadata has formed the motivation for a field of research known as *content-based image retrieval* (CBIR), which emerged from the crossroads of the fields of computer vision, databases and information retrieval. After a promising start, CBIR researchers realized that their efforts were being significantly hampered by what became known as the 'semantic gap', which refers to the inability of a machine to fully understand and interpret images based on automatically extracted low-level visual features, such as predominant texture, color layout or color distributions [3]. The obstacles imposed by such a gap have limited the success of pure CBIR solutions to narrow domains.

Much of current research in CBIR—or the broader field of visual information retrieval (VIR)—is aimed at reducing the semantic gap and incorporating textual information in order to improve the overall quality of the retrieval results [4]. For several years, one of the main obstacles faced by researchers trying to combine visual data and textual metadata was the long-held assumption that manual image annotation is too expensive, subjective, biased, and ultimately, not feasible. This assumption has recently been challenged in many ways, e.g., the availability of Semantic Web-related ontologies [5], the popularity of image labeling games [6], and the willingness of users to annotate, tag, rate, and comment on pictures, enabled by social media sharing sites [7]. The latter aspect, namely the increasing availability of user-generated tags, combined with the successful track record of CBIR within narrow domains, has motivated this work.

However, manually tagging images is an extremely time consuming task. Automatic tagging systems may be able to address this problem by tagging images autonomously as they are uploaded. Yet, one obvious problem is that images may be incorrectly tagged or that important concepts may be skipped entirely. Therefore, fully automatic tagging systems may end up hindering the very processes tagging aims to support. An additional issue is that tags may also have multiple or special meanings within a user group. Intermediate and interactive solutions, however, such as assisted tagging or tag recommendation, can circumvent these problems by suggesting potential tags to the user which may then be manually accepted or rejected, thus achieving a balance between productivity and quality.

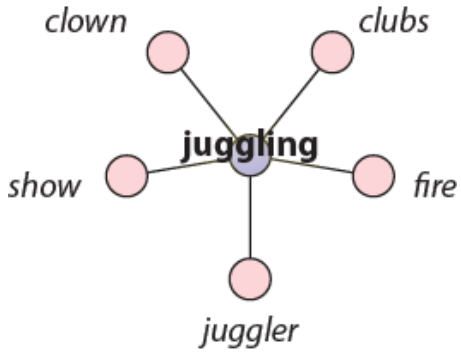
The goal of the research efforts described in this paper is to improve annotation quality and quantity (*i.e.*, increase the number of meaningful tags assigned to an image) by tag recommendation. This was accomplished by developing a tag recommendation system which suggests tags based both on the context of the image and its visual contents. Rather than utilizing synthetic data or tagging vocabularies, it attempts to leverage the "wisdom of the crowds" [8,9] by utilizing existing images and metadata made available by online photo systems, e.g., Flickr. The system is designed to support users throughout the tagging process, to be applicable to a broad range of domains, to be scalable, and to provide realistic performance.

1.1. Use Case

In this section we present a use case to illustrate the proposed approach. We assume that a user uploads a photo of a fire juggling act taken at night with long exposure to Flickr (Figure 1A). We further assume that the user annotates the photo with a single tag: *juggling*. Based on this tag, a number of related tags can be suggested using tag co-occurrence, for example: *clown*, *fire*, *show*, *clubs* and *juggler*. Figure 2 shows an ego-centered network depicting the relations between the tag juggling (in the center) and the suggested related tags surrounding it.

Figure 1. Photos motivating our use case (the first one is the input image with one tag assigned by the user, the others have been previously uploaded by other users and their respective owners have assigned the tags shown below each image [10]).

Figure 2. Ego-centered network of related tags around the tag *juggling*.



Based on the ego-centered network one can assume that the tags *clown*, *fire*, *show*, *clubs* and *juggler* are good candidates to being presented to the user as suggested tags for the input image. However, in typical scenarios, the ego-centered network is not limited to five tags, but may

contain 10–20 times as many, which makes it cumbersome for the user to traverse the list of suggested tags and select the ones she may be interested in. One possible solution is to select the one tag which co-occurs most often with the given tag, but this, too, is subject to errors. In the given example, if we assume that the highest-ranked tag in the suggested tags list is *balls*, we are obviously mistaken.

Therefore we need an alternative ranking approach that allows for visual similarity to be factored into the tag recommendation process, promoting visually similar images (such as the one in Figure 1B) to appear closer to the top of the ranked tag suggestion list, in spite of the relatively low co-occurrence of the tags *juggling* and *fire*. In the architecture proposed in this paper (described in detail in Section 3), we are interested in exploiting visual properties of photos, such as: limited range of hues (mostly yellow to red), large dark areas, and noticeable amount of noise (especially if taken with inexpensive cameras) due to the long exposure involved. This can be done with a principled selection of (pixel-based) feature extraction algorithms and dissimilarity metrics from the field of CBIR and the adoption of the well-known query-by-example (QBE) paradigm, whereby an example image (the one in Figure 1A, in this case) is provided as an example and visually similar images are retrieved from the database. The results of this implicit QBE step, where the example is the image that has just been uploaded, can then be used to strip tags that are assigned to images not visually similar to the initial one. We assume that this leads to a recommendation of better tags in terms of content description. This can result in our example in ranking *fire* and *night* higher, while possibly demoting the tag *balls* from its top position. Consequently, the image in Figure 1B would be ranked higher than the images in Figures 1C and 1D, thereby improving the quality of the tag suggestion process and—perhaps more importantly—allowing the user to retrieve an image that could otherwise have gone undetected (due to the relatively low co-occurrence of the tags *juggling* and *fire*).

1.2. Structure of this Paper

The remainder of the paper is organized as follows: Section 2 presents a broad summary of related work. Section 3 describes the overall picture and proposed architecture for semi-automatic image tagging based on visual features. Section 4 introduces the N-closest photo (NCP) model, promoting its usefulness through an example, explaining how the model is tuned, and discussing the low-level visual features evaluated, selected, and adopted in our solution. Section 5 presents evaluation results of the proposed approach against two different datasets, describes the experimental methodology, and discusses the most relevant results. Finally, Section 6 concludes the paper and Section 7 provides directions for future work.

2. Related Work

Research efforts towards semantically-capable visual information retrieval systems have grown exponentially over the past five years. Some of these efforts are tied to Semantic Web standards, languages and ontologies [11], while others employ keywords in a loose way (not associated with any ontology or folksonomy) [12]. Still others rely on tags (e.g., [6]) and are therefore more closely related to the work proposed in this paper.

Tags assigned by users are often ambiguous, available in several languages or declinations and sometimes not even related to the image content at all [13]. Despite these shortcomings, social tagging

often leads to surprisingly good annotations extracted from a huge amount of annotated content due to the "wisdom of the crowds" effect—the collective knowledge of the user community [8]. Tags may be applied, searched and stored in a very easy fashion; are not restricted to a fixed vocabulary, but instead may be personalized and are just as suitable for small as for large collections [7]. Research on social tagging systems is rather young and therefore continuously expanding into new directions. This section provides an overview of the prominent papers which are relevant for our work in this field.

In [14], Mika analyzes the concept of *folksonomies*, which are the result of social annotations, a network of users, resources and tags, under the assumption that they are social ontologies. Mika presents and discusses an approach for co-assignment analysis in folksonomies, which serves as a basis for part of our work. Network properties in tag co-assignment networks are discussed in [15,16]. In [17], association mining and tag recommendation within social tagging systems is discussed. Hotho *et al.* [18] define a relevance function for retrieval in folksonomies based on PageRank. Tag recommendation for images has also been discussed by Kern *et al.* in [19]. After conducting comprehensive experiments, Kern *et al.* found that for 40% of the images in their test data set from Flickr, correct tags are ranked rather highly.

The approach of Aurnhammer *et al.* [20] is related to our work to the extent that they also postulate that a combination of content-based image features and tags enhances image management. However, while we focus on supporting the annotation process to improve and extend the quality of annotations, in [20] the focus is put on reducing the negative effects of mistaken tags (typos and false tags), synonymy and homonymy for retrieval in image databases.

All in all, a lot of work has been done on auto-tagging lately. Examples are for instance Makadia *et al.* [34], who propose an auto-tagging approach for images based on visual information retrieval. The nearest neighbors in terms of color and texture are determined and labels are transferred from the result set to the query image. Li *et al.* [35] also present a method for auto-tagging employing the vast amount of already tagged photos in the internet. Like in [34] labels from visually similar images are transferred to a query image. In [36] Li *et al.* propose a tag relevance scheme based on visual similarity of tagged images. All these auto-tagging approaches have one problem in common: the semantic gap [4]. With our approach of tag recommendation, where at least one start tag is given, the domain of possible photos is reduced by filtering photos by the given tag(s). Therefore we operate in a small domain, where success to bridge the semantic gap is more likely than in broad and general use cases.

Graham and Caverlee [21] examine the problem of supporting users in the tagging process on a general level. After re-considering the fundamentals of tagging, they explore the prerequisites for implementing tagging in a diverse range of systems outside the traditional strongholds of photography, social networking and bookmarking. Most importantly, such systems should provide high quality recommendations, be adaptable to the individual needs of users, be lightweight enough to be highly usable and take advantage of the collective knowledge of the user community. These principles were taken as axioms for the development of the system described in this paper. Their investigation of related work reaffirms that tagging patterns in large communities stabilize over time, exhibiting clear structures that may be exploited for tag recommendation. Also, similarly to Paolillo and Penumarthy [22], they consider the social aspects but also note that while tagging is a viable concept in the long term, the majority of web content is currently untagged.

Graham and Caverlee [21] have developed an interactive tagging system based on the concept of incorporating user feedback using the general information retrieval techniques of term–based, tag-based and tag-collocation relevance feedback. The core of their system may be summarized as follows:

1. Determine the top-*k* most relevant objects with respect to the target object *o*.
2. Extract and rank tags from these objects.
3. Allow the user to accept or reject the extracted tags.
4. Revise the description of *o* and repeat steps 1–4 until a stopping condition is met.
5. Optionally, allow the user to add additional tags that were not suggested by the system.

Graham and Caverlee went on to implement the process in a service-based interactive tagging framework known as *Plurality*. The system allows documents to be tagged interactively using a model based on the well-known *vector space* model [23]. When a document is first submitted, the system recommends tags using the nearest neighbor paradigm operating on tuples of the form *(User, Document, Tag)*, comparable to the structure presented in [24]. The documents themselves are compared using *cosine similarly* and *term frequency—inverse document frequency* (TF-IDF), which have also been borrowed from classical information retrieval. The tags are suggested by weighing them proportionally to how often they were used on similar documents. This methodology also forms the basis for the *N-Closest Photos (NCP)* model described in Section 4.

The second element of the Plurality system is the *associated feedback* model. During development, three different models were examined for giving feedback on the suggested tags:

- *Tag feedback*. After receiving the proposals, users are given the opportunity to classify each tag as *relevant* or *irrelevant* to the document. The system then uses the information to re-retrieve better matching documents to compute (hopefully) better suggestions.
- *Term feedback*. In this model, users rate the terms extracted from the document rather than the resulting tags, once again with the aim of retrieving more relevant documents.
- *Tag collocation (co-occurrence)*. The system exploits the collective intelligence of the user community by finding tags that were used together with tags the user rated as relevant.

Despite the fact that user feedback is not explicitly considered here, the proposed use scenario indirectly exploits both tag feedback and co-occurrence. In our system (as described in more detail in Section 4), users can accept or reject tags (*i.e.* provide tag feedback) after each iteration. These are then used as start tags in following cycles to further refine the process. Tag co-occurrence, on the other hand, is exploited to provide the tagging vocabulary: only tags that co-occur with start tags may be suggested.

Graham and Caverlee evaluated their system by recording the tagging sessions of 200 participants who were required to use the system to tag documents on *Delicious* (http://del.icio.us). Although they found that users working with the tag feedback model consistently required the least steps to tag documents, the tag co-occurrence model maximized the ratio of selected tags to contributed tags.

In [25], Sinha and Jain present a method for utilizing both an image's content and the context surrounding it to extract semantics. The premise is that interpreting high level semantic data is easier when surrounded by sufficient contextual information. The authors subsequently propose a system to

fuse content information with two types of contextual information, namely optical and ontological, and demonstrate its effectiveness in classification and annotation tasks.

The research of Sinha and Jain is relevant to this work as an example of a system that attempts to classify and automatically tag photos by fusing metadata sources. Similar to the evaluation presented here, they used photos obtained from *Flickr*, citing that these are more representative than more homogenous professional datasets, such as the Corel image set [26]. They do, however, note that many of the photos available on *Flickr* contain unreliable or incorrect tags and subsequently filtered photos and tags to obtain their test dataset.

In their work, Sinha and Jain focus on utilizing the optical content layer, using *aperture*, *exposure time*, *ISO* and *focal length* as parameters and defining the *LogLight* metric, which can be used to determine the proximity of various combinations of camera settings and is designed to reflect the amount of ambient light available when the photo was taken.

Using unsupervised clustering on a dataset consisting of the optical context of 30,000 photos, they produced eight representative clusters of settings that can be considered typical. Next they examined a smaller subset of 3,500 tagged photos, determining the most common tags for each cluster. The results can be considered surprising: the clusters contained mutually exclusive dominant tags, indicating that the clustering can expose semantic information. For example, the long exposure cluster included tags such as *night*, *fireworks* and *moon*.

Sinha and Jain support their claim that optical context information may be used to derive higher level contextual information by building a classifier that places photos into one of three location classes (“outdoor day”, “outdoor night” and “indoor”) based on the *LogLight* metric. Interestingly the classifier achieved an accuracy of 87.5% when using optical context alone. The authors then further extended the classifier with thumbnail image-based features such as average color, color histogram, edge histogram and Gabor texture features, yielding a 2% improvement in accuracy. They note that the high degree of accuracy obtained from the simplistic and compact optical context (typically a few bytes and insignificant compared to the image itself) is worth exploiting particularly when compared to the computationally expensive image features.

The authors then turned their attention to the broader task of automatically tagging photos. They considered two datasets: one crawled and filtered from *Flickr* and the other consisting of manually tagged photos. By applying a Bayes network model to the task, they were able to estimate the probability of a tag being applied to a photo in the testing set based on its frequency in the training set and similarity in image features. To demonstrate the applications of contextual information, the authors also integrated optical clustering based on the *LowLight* metric to improve the accuracy of their model. The five most likely tags for each photo were then taken as suggestions and compared to the actual tags in terms of precision and recall. Results were mixed, achieving a precision of 0.22 and recall of 0.35 for the manually tagged dataset and a precision of 0.14 and recall of 0.27 for the *Flickr* dataset.

As a final step, and perhaps of most relevance to this work, they explored the possibility of improving the tagging system by integrating ontologies, or more specifically, by harnessing related tags. This makes sense as tags are rarely applied to photos in an independent fashion. For example, *fireworks* and *night* are likely to occur together. On the other hand this makes the calculations as presented above, which assumes statistical independence, significantly more complex. Using a restricted set of tags from a lexicon structured as a tree, the authors were able to calculate the similarity

of tag pairs and integrate it into their modeling, once again further increasing precision. Sinha and Jain conclude by stating that extracting high quality semantic information from photos will ultimately require as many knowledge sources as possible. The challenge is thus to design a system which is flexible enough to utilize this diverse range of knowledge.

Probably the most directly relevant body of research to the work presented here is that of Wang *et al.* on annotating images by mining image search results [27]. Citing problems with existing query by example image retrieval systems, they propose a system to circumvent the semantic gap by automatically deriving image annotations (additional tags). Wang *et al.*'s experiments demonstrated that the model-free approach can derive annotations of acceptable precision in real time (under a quarter of a second).

Of particular note is also the scale of Wang *et al.*'s investigation. By mining from a variety of sources, they were able to utilize 2.4 million images, roughly an order of magnitude more than the number of images used in our work. They also placed emphasis on performance and scalability, choosing to construct the system using a series of distributed web services. This was also the motivation behind choosing hash codes to internally represent images which can be compared using nothing more than an "AND" operation.

Similar to the system described in this paper, they too require that the target image be submitted with an initial keyword (start tag). Their system is also *model-free* and, acknowledging the general lack of training data, they source images and metadata from web search engines which are then filtered and combined to produce annotations. The authors cite that this approach has three advantages which are also relevant to our design:

- *No training data is required.* Not using supervised learning removes the need for training data and thus solves the problem of obtaining the required quantity of accurate training data.
- *No predefined vocabulary is required.* This also prevents the system from being limited to a particular domain.
- *Exploitation of the web as a data source.* Compared to previous data sources (e.g. the well-known Corel image set), the web offers a much larger amount of image data, some of which is accurately labeled. This ensures that a diverse range of images is available for a given semantic concept and vice versa.

While Wang *et al.* rely on simplistic (albeit efficient) Hamming distance calculations between image hash codes to locate similar images, our work applies more explicit similarity functions to measure distance between features that reflect particular aspects of an image. Furthermore, their system utilizes all of the text-based metadata that surrounds the images and then clusters results using *Search Result Clustering* to derive annotations, whereas only the tags directly associated with the photos retrieved by the system are considered in our work. On the one hand, their approach, based purely on the axiom "data is the king", is even more general as it makes fewer assumptions about the nature of images or their metadata. On the other hand, however, this increases the prevalence of problems related to text ambiguity and understanding. Subsequently, the authors implemented post-processing steps to reject uncertain words by considering maximum cluster size (larger clusters represent more important attributes of the target image) and maximum average member image score (clusters with smaller intra-cluster variance provide greater certainty).

3. Architecture for Semi-automatic Image Tagging and Annotation Based on Visual Features

Social media sharing sites, e.g., Flickr (http://www.flickr.com), Zooomr (http://www.zooomr.com) or Smugmug (http://smugmug.com) to mention but a few, allow users to form communities which reflect their (photo-taking) interests. Many photos are annotated using keywords called *tags*. Tags are chosen by the user and not restricted by a taxonomy or vocabulary. The process of assigning tags to resources is often referred to as *tagging*.

Some examples of common tags assigned to digital photos on the Flickr web site (sourced from Flickr's all time most popular tags) are provided in Figure 3. On the most basic level, tags can be considered as keywords, categories labels or markers. Some systems may introduce more structure into tags and even so called machine readable tags that encode specific information intended to be read by the system itself. Despite the fact that many platforms allow users to enter free text tags (including spaces and punctuation), the work described in this paper only considers clean, case-invariant single word tags (largely comparable with Flickr's clean tags).

Figure 3. Flickr's all-time most popular tags. In this visualization—called *tag cloud*—the font size of the tag visualizes roughly the frequency of a tag.

Most tagging systems allow multiple tags to be assigned to each object. In this work, the set of tags assigned to an object (*i.e.* a photo) is referred to as a *tag set* (see image in Figure 4 and the assigned tags). The Flickr web service, for example, allows users to provide up to 75 tags for each photo. It is important to note that tag sets are not lists—they do not preserve order.

The highly flexible nature of tagging systems allows tags to be used for many different purposes. When applied in a wider sense, tagging leads to emergent structures that serve a particular purpose [13]. For example, the tag *FlickrElite*, despite not providing any explicit meaning, is used on Flickr to indicate that a photo belongs to a group of elite photographers. In their examination of the social issues surrounding tagging, Paolillo and Penumarthy imply that user groups collaboratively tend

to develop a tagging vocabulary to suit their individual needs and as a result, tags from such "folksonomies" [13] often hold little value for users outside this group [22]. Similar phenomena may also be observed in the tagging structures utilized on the Flickr web service. Despite these differences, tags may be classified into the following four broad classes:

(a) Tags relating to direct subject matter of the photo, *i.e.* describing the objects visible in the photo. For example *old town* or *architecture.*

(b) Tags that provide technical information about the photo itself or the camera, such as whether a flash was used or the particular camera model. For example *nikon* or *i500*.

(c) Tags describing the circumstances surrounding the photo or the emotions invoked by it. For example *awesome* or *speechless*.

(d) Additional organizational tags, often pertaining to the perceived quality of the image or used to identify individual users or groups. For example *diamondclassphotographer* or *flickrelite*.

Photos are typically tagged using tags from all four categories, as demonstrated by the example shown in Figure 4, in which the photo is tagged with a total of 25 tags: 12 relate to the content of the image *(bee, flower, clouds, sky, fly, nature, fleur, fleurs, flowers, insect, insect* and *animal*), one concerns the technicalities of the photo (*macro*), eight describe the circumstances and emotions (*buzz, mywinner, anawesomeshot, abigfav, abigfave, aphotoday, eyecatcher* and *bravo*), and the remaining four have an organizational function (*project365, freedp, shieldofexcellence* and *frhwofavs*). Note that a photo's set of tags may also contain synonyms, sometimes in a foreign language, or even pluralizations.

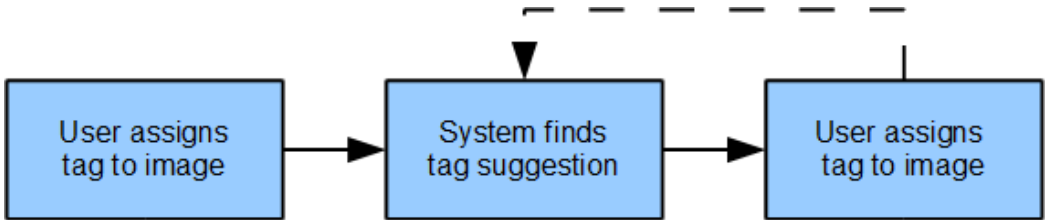
Figure 4. A sample image on Flickr together with its tags (Lift Off by Flickr user aussiegall. Tags: "bee", "flower", "clouds", "sky", "fly", "buzz", "mywinner", "abigfav", "anawesomeshot", "abigfave", "aphotoday", "project365", "eyecatcher", "freedp", "shieldofexcellence", "bravo", "macro", "nature", "fleur", "fleurs", "flowers", "insect", "insect", "animal" and "frhwofavs").

Obviously, the distinction between categories is not always clear. For example, in this instance the tag *macro* could also be seen as a description of the subject matter. In this paper, a distinction is made between *concrete tags* and *noise tags*. The concept of concrete tags is introduced in this work to refer to tags that are directly related to image content (typically classes 1 and 2). Noise tags, on the other hand, refer to tags that hold little inherent meaning that can be applied regardless of the image content (often from classes 3 and 4). The term itself is derived from the concept of noisy metadata introduced in [7]. In this paper, more emphasis is placed on suggesting concrete tags as they tend to be of greater value to end users. Furthermore, it seems reasonable to hypothesize that tagging systems will be less effective in suggesting noise tags due to the fact that they are often unrelated to the image itself and less likely to explicitly co-occur with other tags. Conversely, recommendations based on image features are likely to be more accurate for tags connected with a photo's subject matter than, for example, tags describing its perceived quality.

General Overview of the Proposed Architecture

Figure 5 provides a general overview of the main user actions as well as the tasks to be performed by the proposed system. We assume that the user has already assigned at least one tag to the input image, which is depicted in the first process (leftmost block) in Figure 5. The system then uses those tags to produce a set of related tags (middle block in Figure 5), based on co-occurrence and visual features derived from images annotated with those tags and produces a ranked list of suggested tags to the user (rightmost block in Figure 5).

Figure 5. Overall process of the proposed system. The dashed line indicates the possibility that after a user has selected a suggested tag the system can recommend further tags based on the selection of the user.



4. The N-closest Photos (NCP) Model

Perhaps the simplest tag recommendation model is to suggest tags that frequently co-occur with the existing tags of an image. In our work, we refer to this model as the Statistical Co-Occurrence model or simply the SCO model. Obviously the biggest disadvantage of this approach is that it always suggests the most frequent co-occurring tags regardless of the actual content of the photo. In addition, recommendations may be dominated by noise tags, like *fantastic*, *abigfave* or *flickrdiamond*, as the common denominator of a group of photos, further reducing the usefulness of the method. In this paper we propose extending this model by taking into account content-based image similarity in addition to tag co-occurrence. The approach proposed in this paper can be thought of as a localized version of the SCO model: first similar photos are found and then the most frequently occurring tags within this

group are used to find and rank tag suggestions. This method, referred to in our work as the N-Closest Photos or NCP model, consists of the following steps:

1. *Access a large collection of tagged photos.* An existing collection of existing tagged photos is a prerequisite for the algorithm. Large online photo sharing sites, such as *Flickr*, are a source of such images as they can be accessed free of charge and provide an immense amount of user annotated images.
2. *Locate photos with the current start tag.* The user is required to enter at least one start tag to describe the target photo. Based on this start tag, a group of photos is retrieved. This set of photos tagged with the user specified start tag is called G, consisting of the $|G|$ most relevant photos for the start tag. G is restricted to the most relevant images as the set of all images tagged with the current start tag can easily grow to a size of millions of images.
3. *Find similar photos.* A subset of similar photos should be extracted from G using a content based distance function to compare photos with the target photo. A group of photos $N \subseteq G$ of the $|N|$ images most similar to the input image is selected.
4. *Synthesize tags.* Finally, the tags used within N are combined to select and rank a number of C tags for the target photo. We employ tag frequency weighted by the rank of the photos tagged.

The following example illustrates how the NCP model works in practice. Consider the photo in Figure 7 and assume the photo was submitted with the start tag *beach* to the NCP model. The actual application of the model (using real data) is shown in Figure 8.

Figure 7. A sample image of a beach to be tagged (Coogee Beach by Flickr user laRuth. Tags: *australia*, *nsw*, *2004*, *coogee beach*, *beach*, *sunrise* and *favorites*).

Step 1 in Figure 8 indicates that the photo is submitted with the start tag *beach*. In step 2 a large amount of photos tagged with *beach* is retrieved (set G), which are ranked in step 3 according to their visual similarity to the initial image. Based on this ranking, the set N, being the most similar photos, is created. Tags assigned to photos in N are ranked according to their frequency within N. In step 5 the C tags ranked highest (shown with $C = 8$) are presented as recommendations to the user.

Figure 8. The *N-Closest Photos* (NCP) model in use.

4.1. Parameters of the N-Closest Photos (NCP) Model

The performance of the NCP model is dependent on five main factors:

- **The number of photos considered for content based similarity |*G*|**. Allowing the algorithm to consider a large number of photos has two main advantages: (1) the more photos the algorithm is exposed to, the greater its vocabulary, as it can only recommend tags that have been applied to this set of photos; (2) using more photos increases the likelihood of a direct match with a similar photo. On the other hand, the primary disadvantage of using more photos is the associated increase in processing and downloading time. This may be partially addressed through caching and system design.

- **The number of similar photos used** $|N|$. The size of the close photo group N must be chosen with even greater caution. On one hand, larger values of N result in more tags participating in the combination process. This in turn increases the chance of weak matches occurring with many photos that can collectively influence the tag output stage. On the other hand, larger values of N also increase the chance that a single strong match is overcome by a series of irrelevant weak matches.
- **The criterion used to determine similar photos**. Probably the most complicated aspect of implementing the algorithm is selecting a basis for extracting a subgroup of similar photos. In theory, aside from requiring the photos be ranked via a distance function, any image feature descriptors employed in visual information retrieval may be used. However, the choice of the feature influences accuracy and runtime of the approach.
- **The number of tags synthesized** C. In theory, the algorithm can generate an arbitrary number of tags for the target photo, however choosing a large value of C may not be helpful to the user and lead to lower precision.
- **The method used to combine the tags of the similar photos**. Based on the set N all tags assigned to images of N can be considered for recommendation. However, the ranking function for the tags has critical impact on the C selected tags.

4.2. Low-level Visual Features

In this work we focus on the Color and Edge Directivity Descriptor (CEDD), developed by Chatzichristofis and Boutalis in [28], as it is a relatively new low-level feature that has shown promising results. Combining both color and texture information in a 54 byte feature, the descriptor's compactness, together with the low computational effort involved in its derivation and its retrieval performance, compared to common global feature descriptors, make it suitable for applications involving large numbers images. Essentially, the descriptor uses two separate components to determine color and texture information which are then combined into a single joint histogram.

In the course of their investigation, Chatzichristofis and Boutalis compared the performance of CEDD to that of other well-known descriptors, including color descriptors such as Dominant Color Descriptor (DCD), Scalable Color Descriptor (SCD) and Color Layout Descriptor (CLD), as well as texture descriptors such as Edge Histogram Descriptor (EHD). In addition to being more accurate in terms of Average Normalized Modified Retrieval Rank (ANMRR), it was also shown to be up to an order of magnitude faster.

5. Evaluation

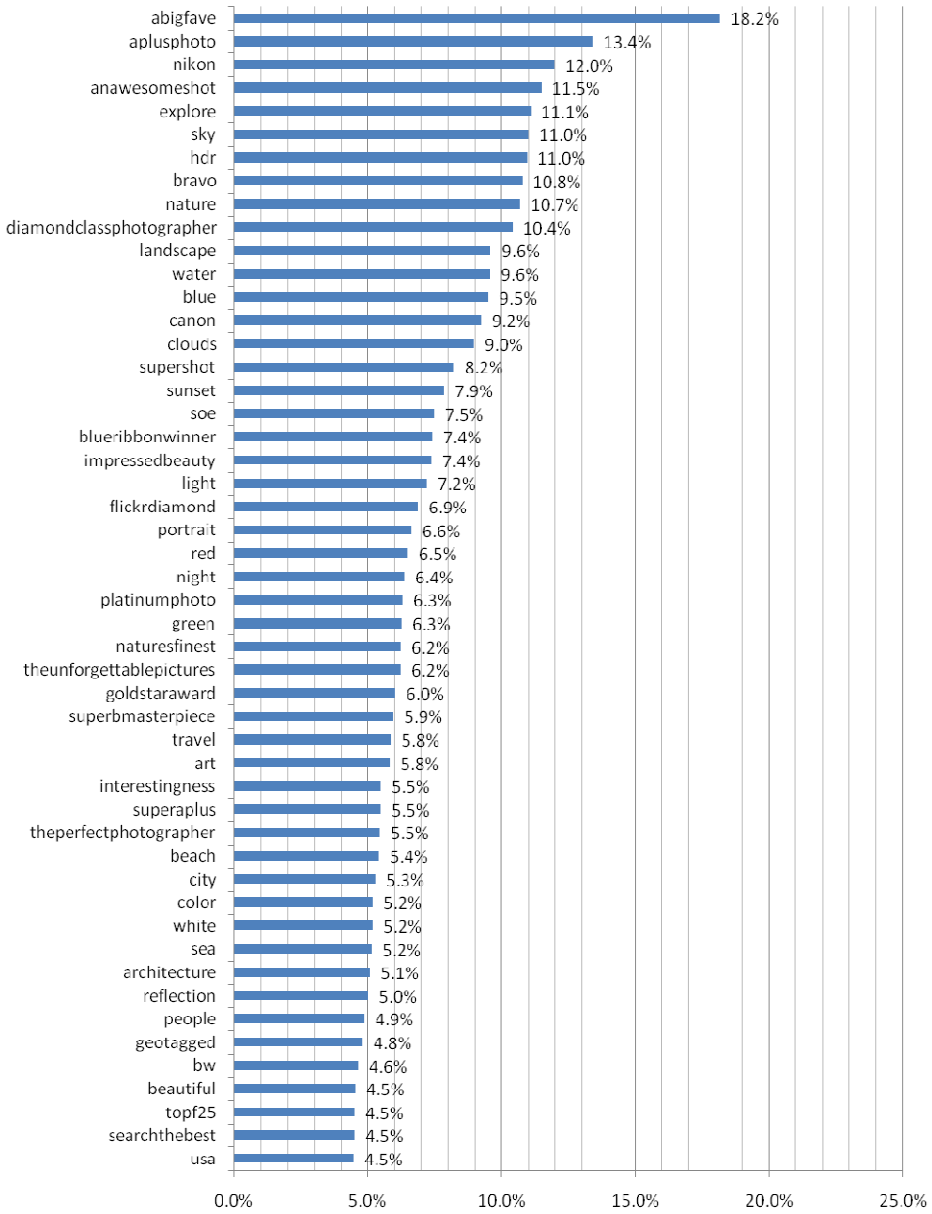
To show the applicability of our system and ensure the relevance for real world applications, we chose to create a dataset based on real Flickr photos. Our basic evaluation strategy was to retrieve a large amount of already tagged photos and use them as gold standard for tag recommendation evaluation. We employed precision and MAP for comparing the statistical co-occurrence based approach (SCO) to our N-closest photo model (NCP). We further investigated the overlap of recommendations to determine if both approaches yielded the same results and we evaluated the qualitative difference in recommendations.

5.1. Dataset

To avoid bias, the set of 144 "all time most popular tags" as determined by Flickr were used as start tags. 400 photos with metadata were collected for each start tag. When locating photos on Flickr, the "most interesting" photos were retrieved first, as determined by Flickr itself. *Interestingness* is a patented relevance function developed by Flickr to describe images "that are beautiful, amazing, moving and striking" [29]. Each photo was then processed to extract their CEDD feature descriptors.

As the existing tag set of each photo (obtained with the photo from *Flickr*) was used to evaluate the effectiveness of the auto-tagging algorithms, it was important that the majority of photos were associated with a predefined minimum number of tags. Photos in the dataset are assigned 26 tags in average with a standard deviation of 15 tags, with 99% of photos having five or more tags, 94% with 10 or more and 59% with 20 or more. Notably, Wang *et al.* found in their evaluation that photos were typically annotated with 19.6 words [27], which is slightly less than in our data set.

Figure 10. The 50 most frequently used tags in our data set.



The 50 most frequent tags of the dataset are given together with their frequencies in Figure 10. A quick examination reveals that almost half of these tags could be regarded as tags not describing the actual content, like *abigfave*, *plusfoto*, *nikon*, *anawesomeshot*, *hdr*, *etc.*

5.2. Methodology

The two most obvious evaluation criteria are *what proportion of the suggested tags is appropriate for the image* and *what proportion of the possible tags for an image was suggested*. These can be likened to the standard criteria of information retrieval, namely *precision* and *recall*. In this case *precision* assesses the number of correct tags found, while *recall* determines the proportion of total tags identified. Formally, for a set of suggested tags *S* and ground-truth valid tags *G*, they are defined as follows:

$$Precision = \frac{|S \cap G|}{|S|}$$

$$Recall = \frac{|S \cap G|}{|G|}$$

One limitation of precision and recall in this scenario is that neither reflects the rank of the retrieved tags, *i.e.* a correct tag last in the suggestions list scores just as highly as a correct tag in first place. This is in direct conflict with the project's use scenario, as users tend to only have a limited amount of time and attention span, meaning that they are more likely to only look at the start of the list. This evaluation utilizes average precision, a variant of precision adapted from the information retrieval domain that places emphasis on retrieving correct tags early [30]. Essentially, it is an average of precisions calculated after each tag is suggested:

$$Average\ Precision = \sum_{i=1}^{N}\left(\frac{1}{i}\sum_{j=1}^{i} R(t_j)\right)$$

$$With\ R(t) = \begin{cases} 1, & \text{if tag t is relevant to the image, contained in G} \\ 0, & \text{if tag t is not relevant to the image, not contained in G} \end{cases}$$

An overview of an algorithm's general performance may be obtained by calculating the *Mean Average Precision* (MAP), *i.e.* the average of the average precision calculated for individual photos.

Generally speaking, MAP was calculated for suggestions for each single start tag available in the data set. The result MAP values were averaged to give an indication of algorithm performance. The algorithm under test was required to suggest a certain number of tags for each photo belonging to the given start tag, with average precision being calculated using the user assigned tags as ground-truth valid tags. In the case of the NCP model, the algorithm was able to utilize information from all photos belonging to the start tag with the exception of the one being tagged. This is commonly referred to as *leave-one-out cross-validation*, a special case of *k-fold cross-validation* [31]. Furthermore, each photo under test was required to have at least C tags.

5.3. Recommendation Performance Results

The feasibility of fixing algorithm parameters, namely C, the number of tag suggestions generated and, in the case of the NCP model, N, was investigated with the intention of simplifying the subsequent tests. Furthermore, this allowed the effect of these parameters to be ascertained experimentally. To limit the computation time involved, a smaller subset of 40 start tags was randomly selected and used during these tests. Other parameters, such as the method of combining tags were held constant.

Figure 11. MAP values for different C with the SCO model with a fixed |N| = 50.

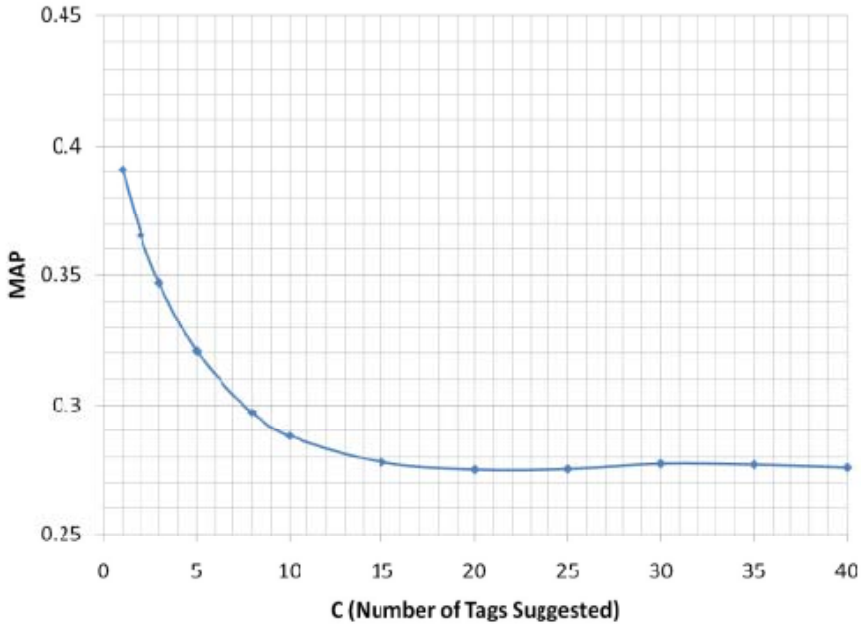


Figure 12. Effect of N on MAP in the NCP model with different values of C.

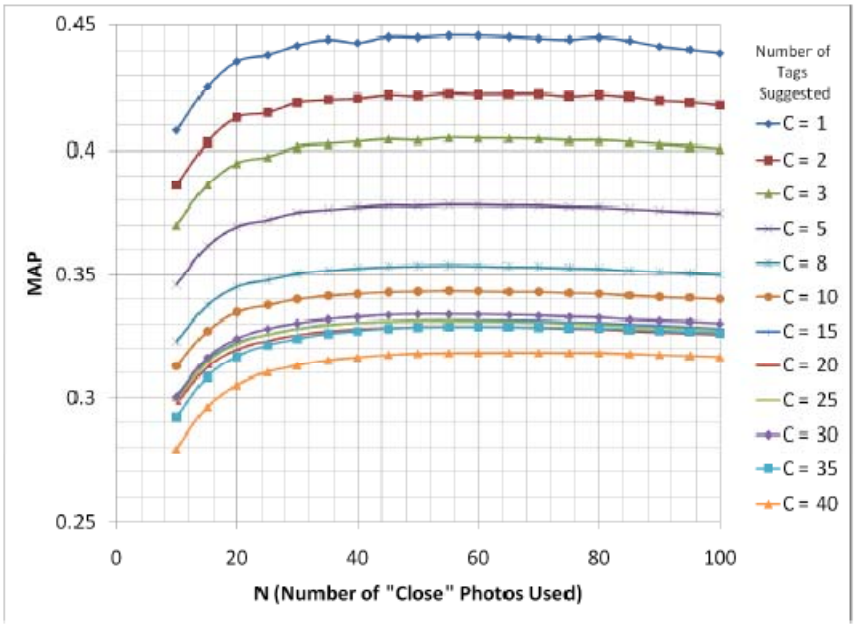


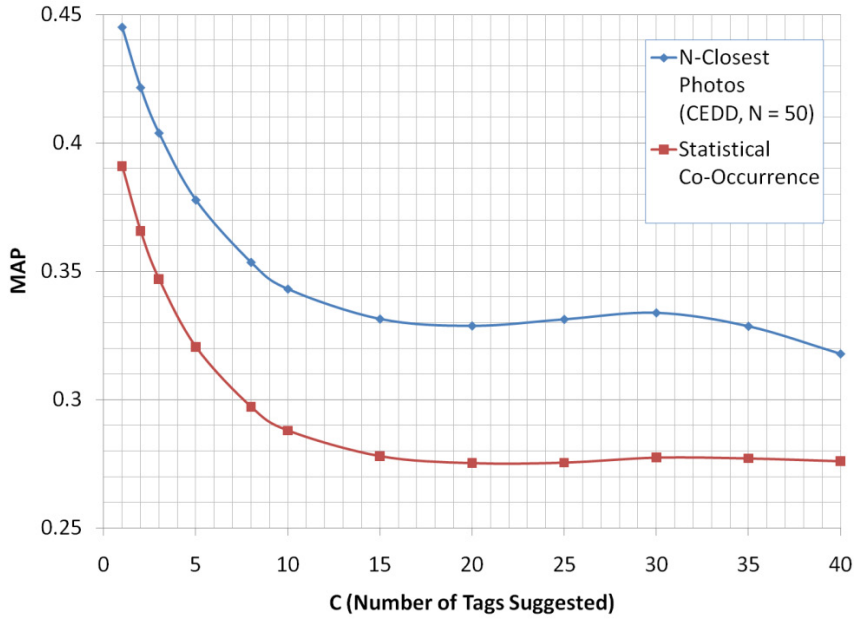
Figure 11 presents the MAP values of the SCO model for a selection of values of C between 1 and 40 with a fixed |*N*| = 50. Broadly speaking, as C increases, MAP decreases. Larger values of C, however, indicate that MAP can be expected to stabilize. Figure 12 shows the MAP of the NCP model under influence of different values for C and N. From the results shown there we can conclude:

- All variants of the NCP model perform well with a size of *N* between 40 and 60. As larger *N* results in increased computation effort, |*N*| should be chosen to be smaller when larger values offer little increase in performance. As a result, |*N*| was set to 50 for the rest of the evaluation.
- The shape of the MAP curve is essentially the same regardless of C. This uniformity leads to the hypothesis that C and N can be set independently.

The degradation in performance resulting from increasing C is progressively smaller as C increases. MAP appears to stabilize for values of C over 15.

Based on the optimal parameters *C* and |*N*| determined from the tests we aim to test the relative performance of the NCP model against the base line SCO model in terms of mean average precision. Figure 13 shows the MAP values of the NCP and the SCO approach. The NCP model performs better than the SCO. However, the algorithms tested delivered MAP values within 3% of one another meaning that improvement offered is marginal.

Figure 13. Comparison of the NCP model with |N| = 50 to the SCO model with different values of C.



5.4. Independence of the Tag Suggestion Models

In addition to the evaluation described in Section 5.3, the approach was also tested upon the redundancy of tags suggested by both models. Although this is a difficult question to answer objectively, we consider the independence of the two models or the overlap of tag suggestions to see if both models recommend the same tags or different ones. The tags suggested by each model for each photo were further analyzed by considering the average sizes of the following four subsets:

- The number of tags suggested by both models
- The number of correct tags suggested by both models
- The number of correct tags suggested by the NCP variant
- The number of correct tags suggested by the SCO model

Table 1 gives the results of this evaluation. In average, 2.53 out of the five tags suggested were the same while the average overlap of correct suggested tags is 0.93. Importantly, the difference between the two models becomes significant here: the NCP model suggests 0.71 extra correct tags per photo ("Average Suggested Tags Correct: N-Closest Photos"—"Average Correct Suggested Tag Overlap") compared to the SCO model.

Table 1. Average correct and overlap set sizes between NCP and SCO suggestions with 95% confidence intervals.

Average Suggested Tag Overlap	2.53 ± 0.10
Average Correct Suggested Tag Overlap	0.93 ± 0.07
Average Suggested Tags Correct: N-Closest Photos	1.64 ± 0.07
Average Suggested Tags Correct: Statistical Co-Occurrence	1.36 ± 0.07

Taking a qualitative look at the concrete tags suggested by both models, we experienced that the suggestions of the NCP model are more related to the visual content (e.g. sun, red, clouds, sea, beach) than the suggestions of the statistical model and therefore more descriptive for the scenes actually depicted on the images.

6. Conclusions

In this paper the NCP model—a novel approach for tag recommendation for digital photo tagging—has been presented. The NCP model incorporates content based similarity between digital photos and achieves slightly higher MAP values for tag recommendation than a typical statistical approach based solely on tag co-occurrence. However experiments with the overlap of the recommendations provided by both approaches have shown that the NCP model yields a reasonable amount of extra information quantified in average to 0.71 additionally recommended correct tags (out of a set of 5 recommended tags). So going back to the question posed in the title of this paper… Can Global Visual Features Improve Tag Recommendation for Image Annotation? The answer is yes. The work described in this paper shows how it can be done and what results are to be expected for the trade off of extra computational effort by incorporating content based similarity.

However, runtime complexity is—in theory—not that high. A search based on a tag, employing inverted files, is sub linear, content based search is linear in time, but can be optimized by (i) reducing the number of images to rank, which is already done in our approach, or (ii) by employing advanced methods for indexing based on low level features. The main practical issue, in terms of runtime, is the download and indexing of images, which includes the extraction of low level features.

7. Future Work

A wide range of future work can be expected in the area, in particular focusing on improving the tag suggestion algorithms and expanding the scope of the evaluation:

Recursive tag suggestion. The original use scenario presented earlier in this article includes the possibility that users can resubmit the accepted tags as start tags to recursively improve suggestion. Citing the opinion of Wang *et al.* in [27], given enough existing images and metadata, it seems likely that recursive tag suggestion would offer sizable improvements in precision. Although the concrete value of this technique would be exceptionally difficult to measure given the exponential growth in the number of photos required, it could be offered as part of a wider evaluation.

Broader evaluation. Directly involving users in the evaluation of tag suggestions, for example by recording their interactions with a system offering the use scenario of Section 1.3, would enhance the significance of the evaluation a lot as the proposed system is an interactive one. Furthermore, the serendipity of suggestions would hopefully lead to an improvement in the quality of tag metadata in general and thus the quality of recommendations.

The integration of advanced knowledge engineering aspects. Clustering (unsupervised learning) and support vector machines (supervised learning) are potential techniques for tag modeling and might be used to create more intelligent recommendations.

New image feature descriptors. Research in the area of CBIR is constantly producing new image feature descriptors, many of which are well suited for use in tag suggestion algorithms. Also local image features have not yet been considered for the NCP model and might lead to significant enhancements in recommendation quality.

Consideration of user intentions. Lately there has been some work done on user motivations for tagging [32,33]. The motivation or intention of users might influence the relevance of tag recommendations for users. Therefore, a next step would be to integrate not only statistical means and content based means, but also user intentions to deliver the most relevant tags to specific users.

References and Notes

1. Graham, J. Flickr rules in photo sharing, as video tiptoes. USA Today, 7 May 2008. Available online: http://www.usatoday.com/tech/products/services/2008-05-06-tech-flickr_N.htm (accessed on 23 August 2010).
2. Flickr Blog: 3 Billion! Available online: http://blog.flickr.net/en/2008/11/03/3-billion (accessed on 23 August 2010).
3. Smeulders, A.; Worring, M.; Santini, S.; Gupta, A.; Jain, R. Content-based image retrieval at the end of the early years. *IEEE Trans. Patt. Anal. Mach. Int.* **2000**, *22*, 1349–1380.
4. Datta, R.; Joshi, D.; Li, J.; Wang, J.Z. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.* **2008**, *40*, 1–60.
5. Troncy, R.; Van Ossenbruggen, J.; Pan, J.Z.; Stamou, G. *Image Annotation on the Semantic Web*; W3C Incubator Group Report 14 August 2007; World Wide Web Consortium: Cambridge, MA,USA; Sophia Antipolis, France; Tokyo, Japan, 2007.
6. Von Ahn, L. Games with a Purpose. *Computer* **2006**, *39*, 92–94.

7. Furnas, G.W.; Fake, C.; Von Ahn, L.; Schachter, J.; Golder, S.; Fox, K.; Davis, M.; Marlow, C.; Naaman, M. Why do tagging systems work? In Proceedings of CHI '06: Conference on Human Factors in Computing Systems, Montreal, Canada, 22–27 April 2006; ACM Press: New York, NY, USA, 2006; pp. 36–39.
8. Guy, M.; Tonkin, E. Folksonomies—Tidying up Tags? *D-Lib Magazine* **2006**, *12*, doi:10.1045/january2006-guy. Available online: http://www.dlib.org/dlib/january06/guy/01guy.html (accessed on 23 August 2010)
9. Surowiecki, J. *The Wisdom of the Crowds*; Vintage/Anchor: New York, NY, USA, 2004.
10. Images are Licensed under Creative Commons by Flickr Users (from left to right): *phatcontroller*, *monkeypuzzle*, *si* and *mathiasl.*
11. Hyvonen, E.; Styrman, A.; Saarela, S. Ontology-based image retrieval. In Proceedings of XML Finland 2002: Towards the Semantic Web and Web Services, Helsinki, Finland, 21–22 October 2002; pp. 15–27.
12. Rasiwasia, N.; Vasconcelos, N.; Moreno, P. Query by Semantic Example. In Proceedings of CIVR 2006: The 5th International Conference on Image and Video Retrieval, Tempe, AZ, USA, 13–15 July 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 51–60.
13. Golder, S.A.; Huberman, B.A. The Structure of Collaborative Tagging Systems. *J. Inform. Sci.* **2006**, *32*, 2.
14. Mika, P. Ontologies are us: A Unified Model of Social Networks and Semantics. In Proceedings of International Semantic Web Conference, Galway, Ireland, 6–10 November 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 522–536.
15. Cattuto, C.; Schmitz, C.; Baldassarri, A.; Servedio, V.D.P.; Loreto, V.; Hotho, A.; Grahl, M.; Stumme, G. Network Properties of Folksonomies. *AI Commun.* **2007**, *20*, 245–262.
16. Lux, M.; Granitzer, M.; Kern, R. Aspects of Broad Folksonomies. In Proceedings of TIR-07: The 4th International Workshop on Text Information Retrieval, Prague, Czech Republic, 23–24 June 2007; IEEE: Regensburg, Germany, 2007; pp. 283–287.
17. Schmitz, C.; Hotho, A.; Jäschke, R.; Stumme, G. Mining Association Rules in Folksonomies. In Proceedings of IFCS 2006 Conference, Ljubljana, Slovenia, 25–29 July 2006.
18. Hotho, A.; Jäschke, R.; Schmitz, C.; Stumme, G. Information Retrieval in Folksonomies: Search and Ranking. In Proceedings of the 3rd European Semantic Web Conference, Budva, Montenegro, 11–14 June 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 411–426.
19. Kern, R.; Granitzer, M.; Pammer, V. Extending Folksonomies for Image Tagging. In Proceedings of 9th International Workshop on Image Analysis for Multimedia Interactive Services, Klagenfurt, Austria, 7–9 May 2008.
20. Aurnhammer, M.; Hanappe, P.; Steels, L. Integrating Collaborative Tagging and Emerging Semantics for Image Retrieval. In Proceedings of the Collaborative Web Tagging Workshop, Edinburgh, UK, 22–26 May 2006.
21. Graham, R.; Caverlee, J. Exploring Feedback Models in Interactive Tagging. In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, Australia, 9–12 December 2008; IEEE Computer Society: Washington, DC, USA, 2008; pp. 141–147.

22. Paolillo, J.C.; Penumarthy, S. The Social Structure of Tagging Internet Video on del.icio.us. In Proceedings of HICSS '07: The 40th Annual Hawaii International Conference on System Sciences, Big Island, HI, USA, 3–6 January 2007.
23. Salton, G.; Wong, A.; Yang, C.S. A vector space model for automatic indexing. *Commun. ACM* **1975**, *18*, 613–620.
24. Marlow, C.; Naaman, M.; Boyd, D.; Davis, M. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In Proceedings of HYPERTEXT '06: The Seventeenth Conference on Hypertext and Hypermedia, Odense, Denmark, 23–25 August 2006; ACM: New York, NY, USA, 2006.
25. Sinha, P.; Jain, R. Semantics in Digital Photos: A Contenxtual Analysis. In Proceedings of 'ICSC '08: The 2008 IEEE International Conference on Semantic Computing, Santa Clara, CA, USA, 4–7 August 2008; IEEE Computer Society: Washington, DC, USA, pp. 58–65.
26. See http://www.cs.princeton.edu/cass/benchmark/ (accessed on 23 August 2010).
27. Wang, X.-J.; Zhang, L.; Li, X.; Ma, W.-Y. Annotating Images by Mining Image Search Results. *IEEE Trans. Patt. Anal. Mach. Int.* **2008**, *30*, 1919–1932.
28. Chatzichristofis, S.A.; Boutalis, Y.S. CEDD: Color and Edge Directivity Descriptor: A Compact Descriptor for Image Indexing and Retrieval; 'Computer Vision Systems. In Proceedings of 6th International Conference on Computer Vision Systems, Santorini, Greece, 12–15 May 2008.
29. National Institute of Standards and Technology. Appendix: Common Evaluation Measures. In Proceedings the Sixteenth Text REtrieval Conference, Gaithersburg, MA, USA, November 2007.
30. Webb, A.R. *Statistical Pattern Recognition*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2002.
31. See http://www.flickr.com/explore/interesting/ (accessed on 23 August 2010).
32. Nov, O.; Naaman, M.; Ye, C. What drives content tagging: the case of photos on Flickr. In Proceedings of CHI '08: The 26th Annual SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy, 5–10 April 2008; ACM: New York, NY, USA, 2008; pp. 1097–1100.
33. Strohmaier, M.; Koerner, C.; Kern, R. Why do Users Tag? Detecting Users' Motivation for Tagging in Social Tagging Systems. In Proceedings of ICWSM 2010: 4th International AAAI Conference on Weblogs and Social Media, Washington, DC, USA, 23–26 May 2010.
34. Makadia, A.; Pavlovic, V.; Kumar, S. A New Baseline for Image Annotation. In Proceedings of ECCV '08: The 10th European Conference on Computer Vision, Marseille, France, 12–18 October 2008; Springer-Verlag: Berlin/Heidelberg, Germany, 2008; pp. 316–329.
35. Li, X.; Snoek, C.G.M.; Worring, M. Annotating images by harnessing worldwide user-tagged photos. In Proceedings of ICASSP '09: The 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; IEEE Computer Society: Washington, DC, USA, 2009; pp. 3717–3720.
36. Li, X.; Snoek, C.G.; Worring, M. Learning tag relevance by neighbor voting for social image retrieval. In Proceedings of MIR '08: The 1st ACM International Conference on Multimedia Information Retrieval, Vancouver, Canada, 30–31 October 2008; ACM: New York, NY, USA, 2008; pp. 180–187.