A Study into Annotation Ranking Metrics in Geo-Tagged Image Corpora

Mark Hughes[1], Gareth J. F. Jones[2], and Noel E. O'Connor[1]

[1] CLARITY: Centre for Sensor Web Technologies,
Dublin City University, Dublin 9, Ireland
[2] Centre for Next Generation Localisation
Dublin City University, Dublin 9, Ireland
{mhughes,gjones}@computing.dcu.ie, noconnor@eeng.dcu.ie

Abstract. Community contributed datasets are becoming increasingly common in automated image annotation systems. One important issue with community image data is that there is no guarantee that the associated metadata is relevant. A method is required that can accurately rank the semantic relevance of community annotations. This should enable the extracting of relevant subsets from potentially noisy collections of these annotations. Having relevant, non-heterogeneous tags assigned to images should improve community image retrieval systems, such as Flickr, which are based on text retrieval methods. In the literature, the current state of the art approach to ranking the semantic relevance of Flickr tags is based on the widely used tf-idf metric. In the case of datasets containing landmark images, however, this metric is inefficient due to high frequency of common landmark tags within the data set and can be improved upon. In this paper, we present a landmark recognition framework, that provides end-to-end automated recognition and annotation. In our study into automated annotation, we evaluate 5 alternate approaches to tf-idf to rank tag relevance in community contributed landmark image corpora. We carry out a thorough evaluation of each of these ranking metrics and results of this evaluation demonstrate that four of these proposed techniques outperform the current commonly-used tf-idf approach for this task.

Key words: Image Annotation, Landmark Recognition, Tag Relevance

1 Introduction

Web sites that store and organise personal image collections online such as Flickr [1] have very large volumes of personal images in their databases. Flickr currently has over five billion personal photos stored online with an average of 3-5 million images being uploaded daily. Unfortunately, the proliferation of shared

[1] Flickr: www.flickr.com

photographs has outpaced the technology for searching and browsing such collections. With this very large and growing body of information, there is a clear requirement for efficient techniques to structure it and present it to users.

Many consumers, tourists in particular, capture large numbers of images in destinations that they visit, and upon return, share these images online with friends and family. One popular genre of images that are being uploaded to online image repositories, and the genre that we focus on in this paper, are photographs containing famous landmarks from around the world. Due to drawbacks in image classification technology, in most cases it is not possible to automatically classify high level semantic information from these images (such as to label them with the name and location of a landmark) based on image content alone.

Due to technological constraints, for high-level semantic image retrieval queries, retrieval systems are forced to rely on text based retrieval methods based on captions created by users, with little or no formal rules on objectivity or detail. This can lead to retrieval errors (an example of which can be seen in Figure 1) due to homogeneous and subjective captions, and in some cases no caption provided at all. Homogeneous captions such as 'vacation' result in poor reliability of individual items in search, and subjective labels are unlikely to be useful for users other than the captioner.

The average consumer, taking a picture with their digital camera or smartphone generally does not pay much attention to how images are stored, organised and retrieved. They simply want a fast and reliable automated technology that allows them to photograph an image and at a later stage retrieve, view and share that image. They don't wish to spend large amounts of time, in what they regard as the monotonous task of providing textual descriptions for images before uploading them to a web site of their choice. Therefore, an automated approach to this task is desirable. In this paper, we present an automated solution to this problem.

Community datasets are undoubtedly a useful resource for image matching processes as many of them contain manually created metadata describing the content and context of each image. There are however, many problems associated with their use. The main issue with these datasets is the unreliability of the relevance and accuracy of their metadata. This paper explores methods to retrieve subsets of semantically relevant tags with which to annotate a test image, from noisy collections of community data, retrieved through the use of an image recognition framework. In this paper we use a corpus of images that have associated geo-tags representing the location where the image was taken, to improve the performance of our image clustering process.

The paper is split into two main sections, the first of which is a description of an image recognition framework that we implemented to gather a set of candidate image annotations from a community dataset which can be used to annotate a test image. The second section describes a number of approaches which are then evaluated with the aim of selecting a subset of the candidate annotations containing those most semantically relevant to a test image. The paper concludes with a thorough evaluation of each of these approaches.

Fig. 1. *An example of the problems associated with homogeneous tags in text based image retrieval systems. Pictured is the top three ranked results (ranked from left to right) returned from Flickr (24-Jan-2012) when searching for images of the famous landmark, the statue of liberty using the query text: statue of liberty.*

2 Background

In the field of Computer Vision, the detection and description of salient image regions is now relatively mature and several algorithms exist that can detect salient regions and create a highly discriminative feature descriptor to describe the region [10][7]. These feature descriptors can then be applied to find corresponding regions in multiple visually similar images with a high level of accuracy and some invariance to rotation, affine and lighting differences.

Comparing large amounts of these descriptors using a brute force approach is processor intensive and several alternatives have been suggested in the literature to allow for fast comparisons of large numbers of images. Lowe [10] applied an approximate nearest neighbour algorithm based on a kd-tree data structure called best bin first. In similar work, Nister [8] suggested the use of a hierarchical k-means structure, which we adopt in this work as part of an image matching framework to return visually similar images based on a query image. We carry out an evaluation of this approach using our image corpora in Section 3 to first create visually similar clusters of images for each query image.

In a related field, work has been carried out analysing how to best extract representative textual tags from clusters of images in community contributed datasets. Kennedy et al. [1] explored different methods to structure Flickr data, and to extract meaningful patterns from this data. Specifically, they were interested in selecting metadata from image collections that might best describe a geographical region. In similar work [2], they focused these techniques on extracting textual descriptions of geographical features, specifically landmarks, from large collections of Flickr metadata. Tags are clustered based on location, and using a tf-idf approach tags are selected so as to correlate with nearby landmarks.

Ahern et al. [3] employ a tf-idf approach on sets of Flickr tags to create a visualisation of representative tags overlaid on a geographical map. They call this system the 'World Explorer', and it allows users to view unstructured textual tags in a geographically structured manner.

Sigurbjornsson and Van Zwol [6] developed a technique to augment a user defined list of tags with an additional set of related tags extracted from large collections of Flickr imagery. They adopt a co-occurrence methodology retrieving tags that regularly co-occur with each user defined tag within the dataset.

Xirong et al. [4] combine visual information with a tf-idf scoring metric to estimate tag relevance within a dataset of Flickr images. For each test image, they carry out a visual search procedure to find its nearest neighbours visually within the dataset. They show that by calculating co-occurrences of tags within visually similar images, it is possible to estimate relevant tags for a query image over using text based methods alone with a higher probability.

Most of the approaches to date have focused on variations of text-retrieval based models using a tf-idf scoring approach to choose relevant representative tags from a cluster of metadata [5]. We have found that tf-idf is not an optimal ranking metric when dealing with a corpus of landmark images due to high repetition of well-known landmark terms. Therefore, the aim of our work is to improve upon tf-idf, by analysing alternative statistical methods to select semantically relevant sets of tags for a test image.

3 Landmark Identification

In order to carry out an investigation into automated annotation, firstly a method to extract a set of candidate tags for a test image must be applied. In this paper we implement an image recognition framework using a collection of geotagged images from Flickr as a training corpus. This framework analyses a test image and retrieves relevant images from the corpus based on visual similarity. The short textual annotations that accompany these retrieved images (denoted as tags) are then considered to be a list of candidate tags with which we aim to select semantically relevant subsets to use as annotations that describe our test image.

For the purposes of this work, we use training and test corpora consisting of images containing commonly photographed landmark images. The focus is on landmarks due to the significant contribution that they make to a large scale public photo repository such as Flickr (eg. Flickr search for Eiffel Tower returns over 450,000 images, Flickr search for Empire State returns over 370,000 images (June 2011)). Landmarks also tend to have a unique visual appearance that leads to high discrimination values between different landmarks.

3.1 Image Corpora

It was desired to create a dataset of geo-tagged imagery that covered an entire metropolitan region of a large city. The city of Paris was chosen, mainly because in certain regions within the city there is a high distribution of landmarks. Additionally, the Parisian region is one of the most densely populated regions that is represented on Flickr with regards to geo-tagged photographs (490,000 in Paris region as of June 2011).

Our training corpus of geo-tagged images was harvested using the publicly available Flickr API [1]. When using the Flickr API, users can provide a text query

[1] Flickr: www.flickr.com

which is used by the Flickr system to return images relevant to that query. To return possible landmark images, the Flickr system was queried with a list of generic words that might indicate a landmark is present in an image, such as landmark, church, bridge, building, facade etc..

To filter out non-landmark images from the corpus, an approach based on the use of stop words was adopted. To build the stop word list, an image set collected from Flickr consisting of 1000 images was manually inspected and classified as containing a large landmark. This set was labelled as S_1. A further set of 1000 images that did not contain a large landmark, but rather depicted an event or different types of objects, people, and animals was also collected and denoted S_2.

For the set S_1, a list of all associated tags was extracted and denoted as T_1. A second list of tags T_2 was created containing all the tags associated with images in S_2. All tags contained in $T_2 \backslash T_1$ were considered possible candidate tags, however the presence of a tag in $T_2 \backslash T_1$ alone is not enough to indicate that the tag would suggest a non-landmark image. It was decided therefore, to select the tags that occurred the highest number of times in T_2 but not T_1. The final set of stop words was selected based on the tag frequency of each possible candidate tag from $T_2 \backslash T_1$. The frequency was calculated using the following formula:

$$tf_i = \frac{t_i}{|T_2 \backslash T_1|}$$

where t_i is the number of occurrences of the tag i in the list $T_2 \backslash T_1$. If the term frequency was above a threshold of .005 (roughly translating to a frequency of 10), the tag was marked as a candidate tag.

Any image within our corpus containing one of these candidate tags was filtered out. In total, we downloaded just under 200,000 geo-tagged images from Flickr in the Paris region from which over 100,000 were filtered out using this approach, leaving a final training corpus consisting of 90,968 images. From informal empirical inspection this tag filtering approach is quite effective, with the majority of images in our dataset depicting a place or landmark.

3.2 Spatial Filtering

Spatial information provides a useful method for filtering the search space for image retrieval processes, particularly when searching for images containing landmarks as they have a fixed location. The first stage of our image matching framework is to filter out non-candidate image matches based on spatial information. While all of the images within our corpus have associated geo-tags, there is no guarantee as to their accuracy. It is therefore not known in advance what level of spatial filtering will be optimal to ensure the best balance between precision and recall. In this section we carry out some experimentation to ascertain the optimal spatial filtering parameter for our image matching process.

There has been some work in recent years evaluating geo-tag accuracies such as work carried out by Girardin [11] and Hollenstein [12]. Both of these eval-

uations however were based on statistical information without any manual inspection and therefore can only be considered as estimates. Inspired by the work of Hollenstein, we carried out a detailed manual analysis was carried out on a subset of the images contained within the corpus to provide a reliable and accurate measurement of geo-tag precision. A subset comprising of 673 images of 4 landmarks was selected to be analysed. Based on local knowledge of the region, each of these images was estimated to have been photographed within very close proximity (approximately 100 metres) to four different landmarks in Paris (Paris Opera House, Arc De Triomphe, Louvre Pyramid and Pont Neuf Bridge).

The geographical centre point of each of these landmarks was noted, and a bounding box with side lengths of 200 metres was created surrounding the centre. The geo-tags of each of these 673 images were examined, and for each one the distance between the geo-tag value and the associated bounding box was calculated to measure the accuracy of each geo-tag.

The results of this analysis are quite interesting (presented in Table 1), in that they indicate that the geo-tags within this dataset are generally accurate to within a relatively small radius. These results show that the majority of geo-tags that were examined are accurate to within 200 metres from our bounding boxes (over 80%). This is not as accurate as a modern, high end GPS receiver (generally accurate to within 10 metres, depending on the strength of the connection and line of sight), but should be accurate enough to allow for efficient filtering of unwanted images in our image search framework.

A number of spatial queries (673) were used to determine the average percentage of the image corpus that remains after spatial filtering at each distance threshold. We find that the search space starts to grow rapidly once the radius has a value of 500 metres or more. Based on the results in Table 1, therefore, we choose a value of 250 metres to use as our spatial radius in our image classification process (Section 3.3), which represents the best balance between precision and recall.

Distance	No. Of Images	% Images	% Search Space
50m	372	55.2 %	1.2 %
100m	506	75.1 %	2.1 %
200m	545	80.9 %	3.7 %
250m	552	82 %	4.6 %
500m	578	85.8 %	8.8 %
1000m	599	89 %	17.6 %
2000m	625	92.8 %	32.5 %

Table 1. Results describing the number of correct geo-tags for each spatial radius, along with the percentage of correct geo-tags from the subset of those examined

3.3 Image Based Landmark Classification

The first aim of the work described in this paper is based on the automated recognition of landmark images. To achieve this goal, an approach based on the nearest neighbor matching of SURF [7] interest point features is utilised. Brute force matching of interest points is notoriously time consuming and potentially intractable when searching large image corpora, therefore, in this work, an alternative approach is utilized based on approximate nearest neighbour search.

We use the hierarchical vocabulary tree approach proposed by Nister [8] to index large numbers of images features and allow for fast nearest neighbour search. A hierarchical vocabulary tree is a tree structure that is built upon a large visual word vocabulary [9]. In this work, we use a vocabulary size of 250,000. The hierarchical vocabulary tree is a form of a hierarchical k-means algorithm, where the inputs consist of visual words, and the cluster centres outputted from each k-means invocation are used as the pivots of the tree structure.

The algorithm quantises the vocabulary into k smaller subsets at each level using the k-means clustering algorithm on each partition independently. Each quantisation takes place recursively on smaller subsets of data. Instead of the k parameter determining the final number of leaf nodes, k determines the branch factor of the structure.

To classify a test image, firstly, its spatial data is analysed and only images that are located within a geographical radius of 250 metres are retrieved from the image corpus, followed by a hierarchical tree based SURF feature matching process. Each child node within the tree represents a vocabulary feature, and is given an identification number. SURF features are extracted from the test image and propagated down the tree structure, each feature being assigned an ID based on its path down the tree. This list of IDs is then compared against the list of IDs associated with all retrieved corpus images and identical IDs correspond as a match. Corpus images are then ranked based on the number of correspondences, and if that number is above a threshold the corpus image is considered a match.

A re-ranking procedure was then carried out using brute force SURF feature matching based on the distance ratio measure using a ratio value as .7. This was carried out as a confirmation stage to eliminate false positives. As point to point matching is an expensive process, the SURF re-ranking process was carried out only on the top images returned from the vocabulary tree that had a number of tree-based correspondences above a threshold k where we evaluated 4 values for k (5, 15, 25 and 35).

3.4 Landmark Image Recognition Evaluation

To test the effectiveness of the landmark recognition system, a test set was created using the Flickr API. This test collection consisted of 1000 images containing landmarks as the main object within the image, photographed within the Parisian region and not contained within the training set.

The precision of the system was evaluated using 4 separate metrics: Precision, precision calculated over the top 3 images (Precision(3)), top 5 images(Precision(5)) and top 10 images (Precision(10)). We test the precision of

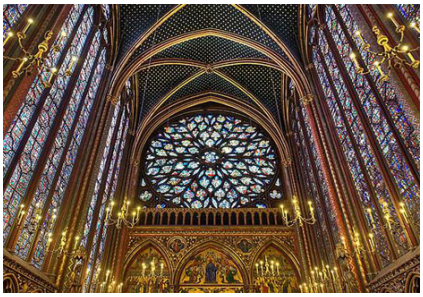


Fig. 2. *Landmark Recognition example: Rose Window in Notre Dame Cathedral*

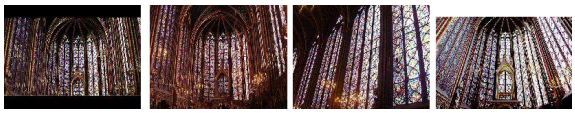


Fig. 3. *The top 4 results retrieved for this image using just the vocabulary tree without SURF re-ranking. While all top ranking images are located within the same structure, they do not contain the same part of the structure present in the query image*

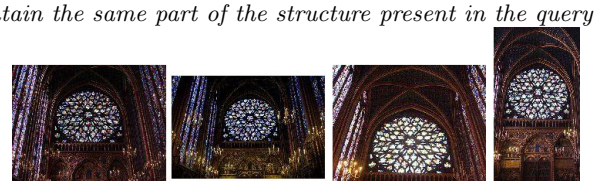


Fig. 4. *The top 4 results retrieved for this image using the vocabulary tree followed by SURF re-ranking. As can be seen in the Figure, the SURF re-ranking significantly improves the accuracy of the object matching retrieving images containing the specific window in the query image*

the system using two approaches, with and without a SURF re-ranking process. We also analyse a number of values for threshold t which determines the number of tree correspondences is required for a corpus image to be considered a match (t=5, 15, 25 and 35).

From the results in Table 3, it is evident that the object matching approach performs with a very high level of precision. The spatial filtering process ensures that the search space is significantly reduced without which the precision would be expected to fall. As expected, the SURF re-ranking process adds a significant improvement over using tree corresponding matches alone. An example of this can be seen in Figures 2, 3 and 4.

4 Tag Selection Schemes

For each test image that we process, the image recognition process will return a large set of tags that are associated with each matched image from our corpus.

Threshold t	5	15	25	35
Precision(Overall)	0.319	0.514	0.612	0.697
Precision(3)	0.695	0.778	0.822	0.851
Precision(5)	0.595	0.711	0.774	0.814
Precision(10)	0.493	0.647	.0.720	0.776

Table 2. Classification results: Hierarchical Vocabulary Tree. The threshold represents the number of tree correspondences required in order for a corpus image to be considered a match

Threshold t	5	15	25	35
Precision(Overall)	0.875	0.936	0.925	0.937
Precision(3)	0.961	0.989	0.988	0.996
Precision(5)	0.934	0.978	0.981	0.984
Precision(10)	0.905	0.955	.0.968	0.970

Table 3. Classification Results: Hierarchical Vocabulary Tree with SURF Correspondence Re-Ranking

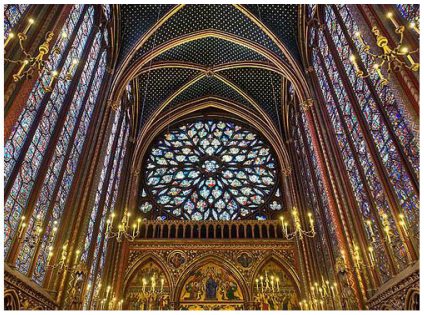
We call this set of tags a result set. There is a significant challenge in retrieving semantically relevant annotations from these result sets, as Flickr tags are notoriously noisy and much of the data is heterogeneous and semantically non relevant. An example of this can be seen in Figure 5. Homogeneous captions are observed to be a common occurrence where a user uploading a large number of images will use the same caption to describe the whole set, which creates obvious problems. The goal is to create a method that will optimally select semantically relevant tags for an image that replicate those that might be selected by a human annotator.

4.1 TF-IDF Tag Selection

Following previous work [3], [4], [1], [5], a method based on the term frequency - inverse document frequency' (tf-idf) approach was implemented and used as a baseline to evaluate all other proposed tag selection approaches. One important measurement in determining the importance of a candidate tag is its level of 'uniqueness' or 'specificity' across the entire corpus. Thus a weighting component is that rewards rarity across the collection is attractive. This is the role of the inverse document frequency (idf). The document frequency of a tag t is defined as the number of images within the corpus that contain t. To scale the weight of the document frequency, an inverse document frequency of a tag is defined as

$$idf_t = \log \frac{N}{df_i}$$

where df_i is the document frequency of the tag i and N is the total number of images within the corpus. Similarly we would wish to reward a tag which appears multiple times for the same item, since it is likely to be an important descriptor for this item, i.e. the term frequency (tf) as defined in section 4.3.



2008	travel	louvre	paris621pams
2009	capital	buttress	arrondisement
toureiffel	city	church	tourist
gothic	military	rosewindow	vacation
otw	gioconda	arcdutriumph	winter

Fig. 5. *A randomly selected subset of candidate tags retrieved by our image recognition framework for the test image shown. As can be seen from the figure, a large percentage of the tags are heterogeneous and of little semantic value to the test image.*

The tf-idf metric is a combination of the idf and tf term is simply formulated as:

$$tf\text{-}idf_t = tf_t \times idf_t$$

In our baseline system each tag within an image result set was assigned a tf-idf score using this metric, and tags were ranked in a descending order with the top k tags selected as the most representative or relevant.

4.2 Proposed Approaches

In this paper we implement and evaluate a number of alternate tag selection schemes to the commonly adopted tf-idf approach. From analysing the structure of the data, five different types of selection schemes were identified:

Tag Selection Based on Term Frequency The first approach evaluated is based on selecting the tag with the highest term frequency score within a result set. Term frequency (TF) is calculated by the number of times a tag appears within a result set, divided by the total number of images within the result set. Tags were ranked based on descending term frequency scores, which essentially corresponds to the terms with majority representation within a result set at the top of the ranking.

Tag Selection Based on Image Similarity Rankings Each result set is ranked based on visual similarity to the query image, with the highest ranking images having the highest number of SURF correspondences. It would seem logical to analyse whether this visual relationship with an image corresponds to contextual similarity within the associated tags. The higher the rank of an image, the more likely it is that the image is a correct match. An incorrectly matched image is more likely to contain irrelevant tags, therefore it seems plausible that higher ranked images have a higher probability of containing relevant tags. To evaluate this hypothesis, a tag selection scheme based on the ranked position of each matched image was implemented. The higher the rank of an image, the larger the weight associated with its corresponding tags. Two weighting schemes were implemented, both of which were based on a mixture of tag frequency within a result set and image ranking.

The first scheme places a large importance on a small number of high ranked images, while the weight associated with images lower down the ranked list is decremented significantly, to such an extent that the lowest ranked images are effectively deemed irrelevant. The score assigned to each tag t is calculated as follows:

$$score(t) = tf_i \times \sum_{i}^{n} w_i \qquad \text{where} \qquad w_i = \frac{1}{r}$$

where r is the rank of the image i in each result set.

The second ranking based scheme provides a more balanced weight across all ranked images. The weight associated with lower ranked images is decremented more slowly. This scheme is formulated as:

$$Score(t) = tf_i \times \sum_{i}^{n} w_i \qquad \text{where} \qquad w_i = 1 - \frac{r}{q}$$

where r is the rank of the image i, and q is the total number of images within the ranked result set.

Tag Selection Based on Ranked Term Frequency Using the Flickr interface, when users are prompted to create tags to describe the content of an image, it can be assumed they they will enter the tags that they deem most relevant to the image in descending order. This order is preserved within the data, and therefore can be considered as a ranked list.

It is logical to assume that if there is a high level of correlation between high ranking tags over a result set of images, that these correlated tags could be deemed most relevant semantically. An evaluation was carried out across all top ranking tags within each result result set. Similarly to the ranked image approach, two different ranking schemes were utilised. The first ranking scheme places a large weight on tags that were ranked near the top of the lists. Tags that are ranked at the lower ends of the list are assigned a weight so low that they are effectively disregarded. This ranking scheme can be formulated as:

$$Score(t_j) = \sum_{i}^{n} w_j \qquad \text{where} \qquad w_j = 1 - \frac{1}{r}$$

where n is the total number of images within a result set in which the tag t_j appears and r is the rank of the tag t_j in image i.

The second ranking approach places a more balanced weight distribution across all tag ranking positions. The variation in weights between top ranking and lower ranking tags is smaller than in the first ranking metric. This second approach is formally defined as:

$$Score(t_j) = \sum_{i}^{n} w_j \qquad \text{where} \qquad w_j = 1 - \frac{r}{q}$$

where n is the total number of images within a result set that the tag t_j appears in, r is the rank of the tag t_j in an image i, and q is the total number of tags retrieved for image i.

Tags are then ranked in a descending order based on $Score(t_j)$. The top ranked k tags are then chosen as the most representative tags for the retrieved image result set.

Tag Selection Based on Geographical Distribution Combining the geographical and textual based metadata that accompanies each image within the training corpus, should improve tag selection precision, as not only does a geo-tag have a semantic relationship with an image, it also has a semantic relationship with the associated textual metadata.

By calculating the spatial distribution of a tag throughout the whole corpus, it is hypothesised that it is possible to predict a relevant tag with a higher probability. A tag with a geographical distribution based over a small geographical area is more likely to describe a landmark within that area, rather than a tag with a citywide geographical distribution.

To indicate the geographically diverse distribution of each tag, a metric calculating the standard deviation is utilised. It is formally calculated using the following formula:

$$dev_i = \sqrt{\frac{1}{N} \sum_{i=0}^{N} (x_i - \bar{x})^2}$$

where x_i is the geographical location for an i^{th} instance of a tag and $\bar{x}$ is the mean geographical location of the tag. All standard deviation values are normalised in the range 0 - 1.

The actual score calculated for each tag is a combination of the tag frequency within the image result set and the geographical variation of the tag. This can be formally defined as:

$$score_i = tf_i \times (1 - dev_i)$$

It was found from experimentation that using a weighted value for tf_i performed better. Based on this, two weights were evaluated:

$$score_i = w(tf_i) \times (1 - dev_i)$$

where w is equal to 2 and 4.

Tag Pair Co-Occurrences The final metric that we evaluated based on the co-occurrence of pairs of tags across each result set. It is believed that it is more likely that semantically relevant tags (e.g. eiffel and tower) would be more likely to appear in pairs in multiple sets of tags than generic tags (e.g. holidays and 2007). We calculate a metric that ranks pairs of tags by the number of times they co-occur.

To calculate the co-occurrence metric, we create a co-occurrence matrix M of size $N \times N$ where N is the number of unique tags returned from within a result set of images,a set denoted as T. The value of position ij in M is number of times tags i and j (where $i, j \in T$) appear together in each retrieved image. The top k pairs of co-occurring tags were then extracted from M (where different values for k were evaluated: 1,2,3,4 and 5). These tags were then re-ranked based on their tag frequency within the retrieved result set of images, where a higher frequency represents a higher ranking.

5 Tag Selection Evaluation

A subset of 100 images was randomly selected from our corpus to be used as a test set to analyse retrieved tags from our image matching framework. To evaluate our proposed approaches, a benchmark selection of tags representing the ranked lists of images returned for each image in this subset was created. Tags associated with each image out of each of these ranked results were analysed manually. This benchmark consisted of a total of 602 retrieved images with an average of just under 6 tags per test image, resulting in a total of 3444 tags. Each tag was deemed semantically relevant or irrelevant to the original test image.

Each approach evaluated analysed different numbers (k) of top ranked tags outputted by our proposed approaches, where $k = [1, 2, 3, 4, 5]$. It is believed that for this task a balanced performance between precision and recall is desired. Based on this, the F-Measure metric is seen as the most important in the evaluation stage. The F-Measure is the weighted harmonic mean of precision and recall. A table displaying the overall F-Measure scores for each evaluated approach is displayed in Table 4.

From the results in Table 4, it is evident that the tag selection scheme based on tag pair co-occurrences performed with the most desirable level of precision and recall out of all evaluated approaches. When utilising a value of 3 for k, where k is the number of selected tags to annotate a test image, there is a recall score of 94% an average precision score of 66% which indicates that for every test image,

Ranking Approach	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
TF*IDF	.43	.51	.57	.58	.57
Term Frequency	.51	.58	.63	.65	.65
Image Ranking $(\frac{1}{r})$	.35	.47	.55	.56	.56
Image Ranking $(1-\frac{r}{q})$	.34	.43	.51	.50	.47
Tag Ranking $(\frac{1}{r})$	.22	.43	.50	.52	.56
Tag Ranking $(1-\frac{r}{q})$	.58	.70	.70	.69	.67
Geographical $(w=2)$	.60	.73	.75	.72	.68
Geographical $(w=4)$	.57	.72	.73	.72	.53
Co-Occurrence	**.75**	**.76**	**.76**	**.73**	**.71**

Table 4. The F-Measure results outputted by each of our evaluated approaches

there is on average 2 semantically relevant tags assigned to it. The selection schemes based on geographical distributions also performed well, outperforming the traditional tf-idf approach, as did the tag ranking based metric which would support the hypothesis that users would enter tags that they discern to be more semantically relevant before heterogeneous tags.

All but one of our proposed alternative approaches to the tf-idf method outperform it in this task (Term Frequency, Tag Ranking, Geographical distribution and Tag pair co-occurrences). Interestingly, the addition of the idf metric to create tf-idf, actually hinders performance over using the tf measure alone. It is thought that the tf-idf metric performs poorly in this dataset due to the high distribution of landmarks. A commonly photographed landmark such as 'The Eiffel Tower' will have a high distribution within the dataset and therefore will have a low idf score which will bias the metric against commonly occurring but semantically relevant tags.

Of all the proposed approaches, the metric based on image similarity rankings performed the worst. It is believed that this is due to the high precision demonstrated by the image matching framework in section 3. The images returned by the system tend to be very accurate which in turn would make the metric redundant.

6 Conclusions

In this paper, several methods were proposed to garner semantic knowledge about a test image from a collection of noisy community metadata. The majority of textual tags associated with this community data is heterogeneous, subjective, and bears minimal semantic relevance from an information retrieval perspective to the content of an image.

Due to the poor performance of the tf-idf ranking metric in this task, the aim of this paper was to propose alternative approaches to tf-idf for rank the relevance of Flickr tags within a visually similar result set. The results of this

evaluation were extremely positive, it can be seen all but one of our proposed approaches (Term Frequency, Tag Ranking, Geographical distribution and Tag pair co-occurrences) outperform the state of the art tf-idf method that has widely been used for similar purposes [3] [2] [4], in this task.

Additionally, as part of this work, we also carried out a detailed manual analysis of the accuracy of geo-tags within our dataset and demonstrated that in this domain, they are accurate to within 200 metres over 80% of the time. Based on this information, we could build an image recognition framework that achieved precision scores of over .9.

References

1. L. Kennedy, M. Naaman, S. Ahern, R. Nair, T. Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia* 631–640, 2007
2. L. Kennedy, M. Naaman. Generating diverse and representative image search results for landmarks *WWW '08: Proceeding of the 17th international conference on World Wide Web*297–306, 2008
3. S. Ahern, M. Naaman, R. Nair, J. Yang. World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections *In Proceedings of the Seventh ACM/IEEE-CS Joint Conference on Digital Libraries* 1–10, 2007
4. L. Xirong, C. Snoek, M. Worring. Annotating images by harnessing worldwide user-tagged photos *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing* 3717–3720, 2009
5. A. Mahapatra,X. Xin,Y. Tian, J. Srivastava. Augmenting image processing with social tag mining for landmark recognition *Proceedings of the 17th international conference on Advances in multimedia modeling - Volume Part I* 273–283, 2011
6. B. Sigurbornsson,R. Van Zwol. Flickr tag recommendation based on collective knowledge *WWW '08: Proceeding of the 17th international conference on World Wide Web* 327–336, 2008
7. H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. In *Proceedings of the 9th European Conference on Computer Vision*, pages 404–417, Graz, Austria, 2006.
8. D. Nister, H. Stewenius. Scalable Recognition with a Vocabulary Tree *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on* 2161-2168, 2006
9. J. Sivic, A. Zisserman Video Google: a text retrieval approach to object matching in videos *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on* 1470–1477, 2003
10. D. Lowe Distinctive image features from scale-invariant keypoints *International Journal of Computer Vision* 91–110, 2004
11. F. Girardin, J. Blat. Place this Photo on a Map: A Study of Explicit Disclosure of Location Information *UbiComp 2007*
12. L. Hollenstein Capturing Vernacular Geography from Georeferenced Tags *Masters Thesis, University of Zurich* 2008