Placing Photos with a Multimodal Probability Density Function

Jonathon Hare
jsh2@ecs.soton.ac.uk

Jamie Davies
jagd1g11@ecs.soton.ac.uk

Sina Samangooei
ss@ecs.soton.ac.uk

Paul Lewis
phl@ecs.soton.ac.uk

Electronics and Computer Science, University of Southampton, United Kingdom

ABSTRACT

Knowing the location where a photograph was taken provides us with data that could be useful in a wide spectrum of applications. With the advance of digital cameras, and with many users exchanging their digital cameras for GPS-enabled mobile phones, photographs annotated with geographical locations are becoming ever more present on photo-sharing websites such as Flickr. However there is still a mass of content that is not geotagged, meaning that algorithms for efficient and accurate geographical estimation of an image are needed. This paper presents a general model for effectively using both textual metadata and visual features of photos to automatically place them on a world map with state-of-the-art performance. In addition, we explore how information from user-modelling can be fused with our model, and investigate the effect such modelling has on performance.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Geo-Location Prediction, Geo-Placing, Geo-Localisation Image Analysis, Tag Analysis, Multimodal Analysis

1. INTRODUCTION

Over recent years there has been a steady increase in the amount of geotagged imagery on the web. Modern digital cameras increasingly contain GPS hardware which automatically attempts to tag the location in which a photograph was taken and the global uptake of smartphones with high quality optics has also led to a big increase in the number of geo-tagged images. That being said, the overall proportion of geotagged images still remains relatively low as demonstrated by the following analysis.

ICMR'14 Apr 01-04 2014, Glasgow, United Kingdom

The social photo-sharing site Flickr reported reaching *6 billion images* in August 2011[1]. In February 2009, it was reported that 100 million geotagged images had been reached[2]; at that time, the total number of images was just over 3 billion, so in percentage terms the number of geotagged images was only $\sim 3\%$. Flickr also stores the accuracy of geotagging, and if we look at only the most accurate images (*accuracy = 16*), then when the 46 million image dataset described in Section 4.2 was collected (Mid-late 2012), the Flickr API was reporting that there were just under 65 million images with a geotagging accuracy of 16; assuming 6 billion total images, this corresponds to just over 1% of the images.

Given that so few images are well geotagged, it is interesting to explore whether it is possible to automatically and accurately predict the geolocation of an arbitrary image. Over the past few years, a number of researchers have investigated techniques that attempt to predict the location of an image using various features and datasets. A summary of these can be found in Section 2.1. Additionally, there have been attempts to standardise the evaluation of such systems through the use of standardised datasets and evaluation protocols. One such example of this is the 2013 MediaEval placing task [11], described in Section 2.2.

The primary goal of the 2013 MediaEval placing task was to develop techniques for accurately predicting the geolocation of a set of Flickr images in terms of latitude and longitude. In addition, a secondary goal was to enhance predictions by estimating the error of the predicted location of each image. The task organisers provided a set of about 8.5 million images with metadata and locations for training, and sets of up to 262,000 images without geotags for testing. Additionally, an evaluation protocol was defined to measure the accuracy of prediction and accuracy of the estimated error (in addition to location predictions, participants were also asked to optionally provide error estimates for each prediction in terms of a distance in Kilometres).

In this paper we describe our state-of-the-art approach for multimodal geolocation estimation. As described in Section 3, the motivation for our technique is twofold; we firstly wanted to develop a technique that can operate using either the visual content or the metadata, but which also seamlessly allowed blending of information across modalities and

[1] `http://latimesblogs.latimes.com/technology/2011/08/flickr-reaches-6-billion-photos-uploaded.html`
[2] `http://code.flickr.net/2009/02/04/100000000-geotagged-photos-plus/`

also allowed information from external gazetteers to be incorporated. Secondly, we wanted the technique to be scalable and efficient, with the aim of being able to estimate the position of an image in well under a second using standard desktop hardware.

In Section 4 we quantify the performance of our technique under different configurations by evaluating it using the 2013 MediaEval placing task dataset and demonstrate state-of-the-art performance. In addition, we explore the effect that user-modelling can have on performance, and relate this back to official results from the 2013 MediaEval placing task evaluation.

2. BACKGROUND

A number of researchers have tackled the problem of predicting the geographic position of a photograph in the past few years. A wide overview of the approaches can be found in the recent survey by Luo et al. [21]. In the following paragraphs, we give an overview of a number of techniques that are either directly relevant to our approach, or served as motivators for the design of our approach. In addition, we also describe the 2013 MediaEval placing task, which took place recently, and detail a number of the approaches taken by the participants. The results of these approaches serve as a comparison to the results we present in Section 4 using the same data and methodology.

2.1 Related Work

Perhaps the oldest work looking at geo-placing is that of Hays and Efros [12]. They presented a purely data-driven approach based on finding visually similar images to each query image from a set of 6.4 million geotagged images. Two techniques were used to propagate the geolocation from the similar images: assigning the location of the 1st nearest-neighbour, and by applying mean-shift clustering to find the location of the dominant mode from the locations of the closest 120 similar matching images. Whilst the technique of Hays and Efros is very general, it provides very limited accuracy. One reason for this is that it is incredibly difficult to explicitly place photos from visual features alone, unless they contain some explicit landmark or a unique feature.

Various authors have developed data-driven techniques that use the textual metadata (typically tags) associated with an image in order to estimate locations. In general, these approaches break the surface of the Earth into cells, and use statistical *language models* that predict the likelihood of an image being located in that cell based on the metadata terms. The language models for each cell are learned using a large corpus of geotagged imagery with suitable metadata. Specific examples of such techniques include the work of Serdyukov et al. [24] who use cells from a fixed size grid; Hauff and Houben [10] attempt to overcome the limitations of a fixed regular grid by using disjoint dynamically sized cells; Laere et al. [16] use multiple overlapping sets of disjoint cells at differing resolutions; and, O'Hare and Murdock [22] use a hierarchy of cells from regular grids at different sizes. One disadvantage of these methods is that the cell structure imposes artificial boundaries on data from which the language models are learned. Most models attempt to circumvent this by applying smoothing functions across the cells. In terms of geo-localisation using textual-metadata only, some authors have used structured knowledge and gazetteers of geographical places; of particular relevance to the technique we present in this paper is the approach used by Serdyukov et al. [24] who augment their grid-based language model technique with information from the GeoNames[3] database by specifically boosting the weights of tags that occur in the list of English place names.

In terms of multimodal approaches that take into account both the visual features of the images being geo-located, in addition to their metadata, a number of techniques have been proposed. Kelm et al. [14] used textual information as the primary information source, and fell back to visual matching for instances where tags do not provide useful information. Van Laere et al. [27] proposed a two-step process in which a target cell is first determined using a language model, and then similar images within the cell are determined, and the final location is determined by interpolating from the locations of the similar images. A similar approach is taken by Trevisiol et al. [26], who localise based on the tags if they are present, but fall back to a cascade of different options, which penultimately includes a visual content-based method, if there are no tags. Finally, rather than attempting to predict geo-locations anywhere in the world, the approach of Crandall et al. [5] uses a combination of metadata and visual features to train classifiers for a relatively small set of landmarks from a fixed set of cities.

A number of techniques, including some already mentioned, use different types of contextual metadata. Notably, Crandall et al. found temporal features to be useful, and Hauff and Houben investigated the use of the user's tweets to enrich the textual data available for predicting the location.

2.2 The MediaEval 2013 Placing Task

MediaEval is an international evaluation campaign in which participants take part in various multimedia-related challenges and benchmark their results against the other participants using a standardised evaluation methodology. The 2013 MediaEval Placing Task was dedicated to the geo-localisation of photographs from Flickr [11]. Participants were asked to provide estimated latitudes and longitudes for a set of test images from Flickr, and optionally, to provide an estimate of the *placeability* — an estimate of the error of predicted location (as a distance in kilometres).

A development dataset of 8,539,050 geotagged Flickr images together with metadata including tags, the time uploaded and taken, and the user for each image. Five test sets of different sizes (between 5300 and 262,000 images) following the Russian dolls approach (the larger test sets contain all images of the smaller test sets) were also provided (referred to as *test1..test5*), and participants were asked to use the largest dataset they could process. The test sets did not contain the actual geo-locations during the running of the challenge, although the ground truth together with all the other parts of the dataset has subsequently been publicly released[4].

Participants in the task were allowed to submit up to five different *runs*. The first two runs had specific conditions attached: the first was only allowed to use the provided data (i.e. the provided metadata and any features extracted from the provided images); the second run was only allowed to use visual information. The remaining runs

[3] `http://www.geonames.org`

[4] `http://www.st.ewi.tudelft.nl/~hauff/placingTask2013Data.html`

were open, and participants were free to use any additional information they liked, with the exception that they could not crawl the test images in order to find their geolocations from Flickr. In terms of evaluation, the ground-truth of the test-sets was used to estimate the error in a series of circles of 1,10,100,1000 kilometres; if the Haversine (great circle) distance between the true location and predicted location was within a given circle radius then it was considered to be correctly localised. Additionally the median error (the median of the Haversine distance between the prediction and actual location) was calculated. In order to assess the performance of the placeability estimates, the linear correlation of actual error to predicted error was computed.

2.2.1 Overview of submitted techniques

Seven teams submitted runs to the 2013 placing task. With the exception of our runs, on which this paper builds, the approaches of a selection of other teams are described here. In terms of techniques using only visual information, the dominant approach was to perform a content-based search against the images in the development set and propagate the geolocation of the first nearest neighbour to the query. Kordopatis-Zilos et al. [15] used product-quantised VLAD [13] features from SURF [1] local descriptors to achieve a precision of 0.60%@1km and median error of 6715km on the largest *test5* set. Similarly, Li et al. [18] combined independent searches using CEDD [3] and BIC [25] features and the L1 distance with rank-aggregation to obtain a precision of 0.37%@1km and median error of 6632km on the smaller *test3* set. The best visual-only run [20] used a SURF based Bag of Visual Words (BoVW) based initial search, followed by refinement of results a second search on the result set using a combined colour, edge and texture descriptor, finished with a geo-visual ranking [19]. The performance of this approach was a precision of 2.8%@1km on the *test5* set; the median error was not reported.

The runs that only used metadata were primarily based on breaking the surface of the Earth into cells and building language models from the tags. The best performing of these approaches, by Popescu [23], processed images with *Flickr machine tags* separately as it was found that by correlating machine tags with geographical coordinates it was easy to get very high quality location estimations; only if an image didn't have machine tags was the language model applied. Using the provided development data this technique achieved a precision of 26.0%@1km and median error of 98.8km on the *test5* set. Popescu [23] also explored what would happen to performance if additional training data (a set of 90 million images), coupled with 'geographicity' scoring (estimation of how likely a tag is to represent a unique place) and user modelling (through modelling the dominant location of all photos taken by the query user, excluding a 24 hour window around the query). The result of these additions resulted in a precision of 43.0%@1km and median error of 2.08km (*test5*).

Of the true multimedia techniques, using a mixture of both visual features and metadata, there were a mix of submissions. These included multi-step decision-based techniques that primarily worked on the tags, but fell back to visual information in the absence of tags (precision 10.37%@1km on *test5* [15]; precision 21.2%@1km on *test3* [2], as well as integrated approaches that treated visual features and textual features equally (precision 20.11%@1km on *test3* [18]. Some runs showed slight improvements when combining the visual and metadata features, whereas others showed slight drops. Li et al. [18] had a slight drop in their official results when they combined the tags with visual features over the tags alone, however they also reported a significant performance increase when using combined features on a validation set of images extracted from the development data. They posit that the difference was due to their validation set containing the photos from the same users as their training set (the official test data contains a disjoint set of users from the development data). In Section 4, we show this is indeed likely to be the case and discuss the issue in more detail.

3. OUR APPROACH

The basic idea of our approach is that we estimate a continuous probability density function (PDF) over the surface of the Earth from a number of *features* (described below) extracted from the query image and/or its metadata. The use of a continuous PDF alleviates the problems inherent with the cell-based methods which require the surface be broken into a grid (or even multiple overlapping grids to deal with scale). In addition, we are able to seamlessly unify the inclusion of different modalities of features (i.e. tags and different types of visual information), as well as support the inclusion of external data provided by sources such as GeoNames, or even sources of contextual information such as Twitter.

In order to estimate the PDF, each feature provides a fixed size set of sample *points* (latitude, longitude) which are then combined. The number of points each feature provides is a variable of our system; by using more points from one feature over another we can essentially weight the importance of a feature higher or lower (see Section 3.2). Once we have all the points from all the features, a kernel density estimator can be used to estimate the probability density at any arbitrary geographical position:

$$P(lat, lng) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{lat - lat_i}{h}, \frac{lng - lng_i}{h}) \quad (1)$$

where h is the kernel bandwidth, (lat, lng) is the coordinate at which the probability is being estimated, (lat_i, lng_i) is the coordinate of the i-th sample (from n total samples), and $K(\bullet, \bullet)$ is the kernel function. In this work, we use a uniform kernel:

$$K_{\text{uniform}}(a, b) = \begin{cases} 1 & \text{if } a^2 + b^2 <= 1 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

however, we note that other kernels could be used. Intuitively, in the case of our uniform kernel, the bandwidth essentially controls the radius of the circle, about the coordinate in question, over which sample points are summed to estimate the probability.

By finding the modes of the PDF we can create an estimate of the location of the photograph by selecting the position of the mode with the highest probability. In practice, density estimation and mode-finding can be combined by applying the mean-shift algorithm [4]. Mean-shift is a standard algorithm to efficiently find the modes of a PDF from a set of samples of that PDF (i.e. the features in our case). The only variable of the mean-shift algorithm is the kernel bandwidth of the kernel density estimator. As mentioned in Section 2, mean-shift has been used in the context

of geolocation estimation in the past; Crandall et al [5] used mean-shift to determine landmarks, and Hays and Efros [12] used mean-shift on the results of content-based image search to determine probable locations. Whilst Hays and Efros's approach is similar to ours, it differs in a number of important ways. In particular, whereas they only considered single (high recall/low precision) content-based features, we consider the fusion of multiple features from different modalities. In addition, Hays and Efros used the mean-shift algorithm for coarse-grained location estimation, with a very large kernel bandwidth. In our technique, because of the way we are using features we are able to use a much smaller kernel bandwidth for very fine-grained location estimation.

Once an estimated position has been generated, we fit a univariate Gaussian over the support points of the highest probability mode, in order to estimate the accuracy of the estimated geolocation as a function of the variance of the Gaussian. The support points are simply the samples that were assigned to the highest mode by the mean-shift algorithm.

3.1 Multimodal Features

The role of each feature used in our technique is to provide a set of geographic points in response to a query image, which can then be fed in to the mean-shift algorithm. As mentioned in the introduction, one motivation of our technique is that it should be fast. This has direct implications for the features we use. This is especially true for the visual features where we are doing content based searches against millions of images. As such, the features described below that we have chosen to use in our experiments are designed to be as efficient as possible.

3.1.1 Location Prior

The location prior models the distribution of where in the world photos are likely to have been taken, or more correctly where geotagged photos uploaded to Flickr are likely to have been taken. The location prior has been used in previous research, although with mixed results depending on the localisation technique [22]. In our case, a constant prior feature was built by uniformly sampling 1000 geographical coordinates from the training data.

3.1.2 Tags

Every tag in the query image is associated with the coordinates of the training images in which the tag appeared. If a tag in the query was unseen in the training data, then it contributes no points. Each tag is considered to be an independent feature. The set of geo-coordinates from the matching training data is sub or super sampled to a fixed size (see the discussion on weighting below).

Unlike many of the previously proposed techniques, we do not filter any of the tags. The intuition is that every tag provides some evidence of geolocation, although that evidence may be a geographically diverse set of places. An example of this is the tag 'Cat', and its Spanish/Portuguese equivalent 'Gato'. Clearly photos of cats could occur on most landmasses of the World, however if you look at the data, its clear that the tag 'Cat' is more likely to occur in the English-speaking world, whereas 'Gato' is more likely to be associated with images taken in Spain, Portugal, and Central and South America.

In the worse case, a tag that provided no added value would just mirror the prior, but even this is still imparting some knowledge as it would hint that the photo was unlikely to be taken in certain places (for example the middle of an ocean).

3.1.3 Visual Features

Visual matching is incorporated into the model in a simple manner. Content-based searches are performed to find similar images to the query. Geolocations are then sampled from the images in the results set in proportion to the score of the result computed by the search engine. This means that results with higher scores (i.e. those that are more visually similar) have more samples of their respective geolocation in the final feature, and thus this makes the probability density proportionally higher for the locations of the most similar images.

We've experimented with three models of content-based retrieval; the first two models aimed to produce high-recall, but (relatively) low precision results, whilst the final method aimed to produce very high-precision, but with very low recall. Each model is described in detail below.

PQ-PCA-VLAD. Firstly, we have experimented with an index of compressed colour VLAD features following the overall approach of Jegou et al. [13]. Specifically, we extracted RGB-SIFT descriptors at difference-of-Gaussian keypoints from each of the images, normalised the features by element-wise square-rooting, and computed the VLAD aggregation of the descriptors for each image with a codebook of 64 centroids. This resulted in 24576 dimensional features, which were reduced to 128 dimensions using PCA. The 128-dimensional features were indexed with a product-quantiser (16 products of 256 clusters [i.e. 8 bits/cluster; 128 bits/image]) to enable fast in-memory search of the complete training data using the asymmetric distance computation method [13]. The number of images returned was limited to a fixed number of the most similar.

PQ-CEDD. We also experimented with CEDD features [3]. In order to get reasonable search speed, we indexed the raw CEDD features with a product-quantiser, as with the VLAD features described above. The product quantiser consisted of 18 products with 256 centroids [144 bits/image]. Again, as with our VLAD features, the number of images returned by the search was limited to a fixed size.

LSH-SIFT. High-precision image content search was performed using a variant of the approach we developed in [9]. Difference-of-Gaussian SIFT features were extracted from the images and form the basis for image comparison. To efficiently assess whether features match, Locality Sensitive Hashing is used to create sketches (compact binary strings) from the features. The sketches are produced such that the Hamming distance between sketches approximates the Euclidean distance between the features [7]. As in [7], we choose our sketches to be 128 bits in length, and set the minimum Hamming distance for two sketches to be classed as matching at 3 bits. Rather than explicitly compute Hamming distances between all features, an efficient approximate scheme is used: The 128-bit sketches are partitioned into 4 32-bit strings and represented as 32-bit integers. For a pair of matching sketches there could be at most 3 different bits, so at least one of the pairs of integers from the sketches must be the same. The four integers from each sketch are used as keys in four hash tables and the list of images containing each respective feature is stored as the value.

Table 1: Effect on placing performance of changing how many images are considered by the VLAD-based content search. These results are based on VLAD features alone.

No. Images	10	50	100	150	200	250
Precision@1km	0.33	0.44	0.45	0.47	0.46	0.46

Some types of SIFT features are very common and tend to occur in lots of images, causing large groups of unrelated images to be linked together in the graph. In order to mitigate this, we filter out hash keys associated with many images. In the experiments presented in this paper, any hash with more than 20 images associated is removed.

In order to perform a query, features are extracted and hashed from the query images. The integer hash codes from the query are used to look-up lists of images in the respective hash tables, and the lists are merged into a single list of `<image, score>` pairs, where the score is the number of collisions.

3.2 Feature weighting

The proposed approach naturally allows features to be weighted. All that needs to happen for a feature to be weighted higher (or lower) is for that feature to return more (or less) coordinates for incorporation into the density estimator; this can trivially be performed by super (or sub) sampling. In the experiments presented in this paper we have chosen to leave the vast majority of the exploration of different feature weightings for future work. For most experiments, we've weighted all features equally, and they all provide exactly 1000 coordinates for the density estimation.

We have however experimented with boosting the weight of tags that are likely to represent place names. We used the GeoNames gazetteer to boost the weight of tag features that were likely to belong to a specific geographic location; any textual tag that could be matched against the the GeoNames "name" or "alternate-name" field was boosted by doubling its number of points from 1000 to 2000 (by adding a copy of the list of points to itself). All other features and non-matching tags remained at 1000 points.

3.3 Implementation

The implementation of our methodology was realised in Java using OpenIMAJ[5] [8] and Lucene[6]. Visual feature extraction and indexing, and the hash-table construction for LSH-SIFT, was performed using Hadoop. For performance reasons, we used an approximate mean-shift implementation inspired by the one in scikit-learn[7]. The approximations stem from using a regular grid for determining the seed points from which to seek modes (rather than using all of the actual data points), and using nearest-neighbours to assign data points to modes, rather than actually assigning them to the mode they converge to.

4. EXPERIMENTS AND DISCUSSION

The experiments presented here follow the exact protocol of the 2013 MediaEval placing task, although we have included additional accuracy measurements. All experiments were performed with the largest 262,000 image *test5* set. The aim of the experiments we present here are threefold; specifically, we aim to: quantify and compare the performance of our technique under a number of different conditions; investigate the effect of adding additional training data; and, investigate the effect of implicit user-modelling that occurs when the training data includes photos by the query users.

The new results presented here differ slightly from our official MediaEval runs [6]. The difference is because we have switched to using a sinusoidal projection of the geo-coordinates before applying the mean shift procedure. The sinusoidal projection attempts to better represent any coordinates near the poles as being in a single place, rather than being spread-out in a linear fashion as with mercator-style projections. For comparison, where a new experimental run mimics the settings of an official run we've also included the original official results.

If any combination of features used in an experimental run produces no coordinates for a query image, then the localisation is set to the geographic north pole (90,0). For content-based search with both VLAD and CEDD features, the number of most similar images considered for providing input to the PDF is set to 100, as was the case for our original MediaEval submissions. This is not optimal by any means, as Table 1 shows, although it does work reasonably well. Finally, the kernel bandwidth for the mean-shift clustering was fixed at 0.01° (with a flat kernel); this was based on a loose optimisation of the precision at 1km using a small validation set for the MediaEval runs, and could undoubtedly be set to a better value for different feature combinations, or to optimise a different metric.

4.1 Baseline experiments

The first set of experiments uses only the training data provided in the MediaEval 2013 placing task dataset. The parameters of the runs were loosely optimised using a small validation set to maximise the precision at 1km as this was the official placing task metric; this has a slightly adverse effect on the median error however. The parameters for the system could be optimised to minimise median error instead (this would lead to a larger kernel bandwidth). Table 2 provides results for different feature configurations. If we analyse these results, we can notice a few things of interest.

The first thing to note is that visual features alone perform relatively poorly for exactly predicting locations; this is be be expected as the vast majority of images do not contain recognisable places. The PQ-CEDD feature performs particularly poorly. The LSH-SIFT and PQ-PCA-VLAD features perform much better than the prior at small radii, indicating that a small number of images can be accurately placed. For bigger radii, the LSH-SIFT gives less accuracy than the prior; this is to be expected as the technique provides very high-quality matches, and we would expect that for a very large proportion of the queries there would be no matches, resulting in the default (0,90) position being estimated. The PQ-PCA-VLAD feature tends towards the prior for large radii; again, this is expected as a fixed number of images is retrieved with this technique, and for images that either match poorly or match well to a very small proportion of the top ranked images then the PDF would tend to look very much like the prior. Combining the PQ-PCA-VLAD and LSH-SIFT features improves placing accuracy at small

[5] `http://openimaj.org`
[6] `http://lucene.apache.org`
[7] `http://scikit-learn.org/stable/modules/clustering.html#mean-shift`

Table 2: Results of baseline experiments using only the data provided in the MediaEval 2013 Placing task. Features key: **P**=prior; **T**=tags; **C**=PQ-CEDD; **V**=PQ-PCA-VLAD; **S**=LSH-SIFT. MEx run codes indicate official MediaEval run results.

Run	Features					Percentage Precision at						Median	Error Est.
ID	P	T	C	V	S	0.1km	0.5km	1km	10km	100km	1000km	Error (km)	Correlation
Prior													
B1	✓					0	0.02	0.05	0.58	1.45	13.15	7410.6	0.209
Tag Feature Combinations													
B2		✓				5.13	17.66	23.4	37.95	44.18	55.24	421.1	0.36
ME3	✓	✓				5.44	17.64	22.97	37.42	43.49	56.28	451.9	0.369
B3	✓	✓				5.11	17.61	23.34	37.89	44.09	56.55	408.2	0.348
Visual Feature Combinations													
B4			✓			0	0	0	0.12	0.44	3.19	7691.1	-0.002
B5				✓		0.21	0.37	0.45	1.08	2.02	13.39	6542.6	-0.014
B6					✓	0.3	0.41	0.44	0.65	1.13	7.94	5806	-0.088
B7	✓			✓	✓	0.35	0.49	0.53	0.94	1.75	12.31	6976.5	0.176
ME2	✓		✓		✓	0.3	0.41	0.44	0.71	1.33	9.42	6885.4	0.059
B8	✓		✓		✓	0.31	0.41	0.44	0.72	1.34	9.43	6897.3	0.091
Hybrid feature Combinations													
ME1	✓	✓	✓		✓	5.19	16.04	20.4	31.29	35.82	46.85	1352.9	0.157
B9	✓	✓	✓		✓	4.89	16.1	20.86	32.1	36.74	47.66	1270.7	0.101
B10	✓	✓			✓	4.87	16.04	20.78	31.97	36.61	48.09	1207.6	0.236
B11	✓	✓		✓	✓	4.91	16.15	20.93	32.18	36.92	48.44	1164.6	0.15

Table 3: Results of experiments using the data provided in the MediaEval 2013 Placing task together with the GeoNames gazetteer to boost place names. Features key: **P**=prior; **B**=geonames boosted tags; **C**=PQ-CEDD; **V**=PQ-PCA-VLAD; **S**=LSH-SIFT. MEx run codes indicate official MediaEval run results.

Run	Features					Percentage Precision at						Median	Error Est.
ID	P	B	C	V	S	0.1km	0.5km	1km	10km	100km	1000km	Error (km)	Correlation
X1		✓				4.85	18.25	24.67	41.08	47.83	57.7	194.4	0.368
ME5		✓	✓		✓	5.21	17.93	23.52	38.17	43.89	54.11	540.1	0.041
X2		✓	✓		✓	4.74	17.43	23.32	37.98	43.65	53.87	556.4	0.028
X3	✓	✓		✓	✓	4.76	17.47	23.34	38	43.7	54.49	532.4	0.161

radii, although our results are still short of the technique proposed by Li et al. [20], which achieved a best visual-only accuracy of 2.8% at 1km (compared to our 0.53%). In the future it would be interesting to fuse the geo-visual reranking [19] approach of Li et al. with our visual features.

The tag-based runs perform relatively well. In absolute terms, the best tag-only run at MediaEval achieved a precision of 26% (compared to our best of 23.4%), however this difference could purely be down to what decision is made about placing the images in the testset with no tags (about 13% of the test images). The addition of the prior tends to slightly decrease precision at low radii, but increase it for higher radii; this is also seen in the decreased median error for runs with the prior.

The multimodal hybrid runs all performed slightly worse than the tag-only runs. This is slightly unexpected, as experiments with a validation set taken from the training set indicated that there should be some improvement (Li et al. [18] also noted this), however, the experiments in Section 4.2.1 indicate that this is because the validation set contained images from the same users as the remaining training set. Improved hybrid performance (over that of the tags alone) could likely be obtained by weighting the different features (see the discussion in Section 5).

The final thing to look at is the correlation of our error estimates to the true errors. The visual-only runs show no correlation, however this is largely in part due to the unplaceable images all being placed at the default point, with a fixed error (5800km), which is clearly uncorrelated with the actual position. All the other runs exhibit a positive correlation which is an encouraging indicator that we are able to estimate the error to some extent with the current technique. No other published work has produced error estimates, so it is not currently possible to compare our performance against other work.

4.2 Adding additional data

If we relax the condition that experiments must only use the provided training data, then we have two options: we can use more data (of the same type; i.e. more Flickr images); or we can incorporate a different type of data into the model. In Section 3.2 we described a method for incorporating additional data from GeoNames by weighting tag features that were likely to correspond to places more highly. The results of experiments that apply this weighting are shown in Table 3. These results clearly show that this weighting helps boost precision in both tag-only and multimodal configurations.

In the early autumn of 2012 (around the same time that the placing task dataset was crawled) we crawled our own dataset of over 46 million geotagged Flickr images with a recorded geo-accuracy of 16. In order to assess the performance of our technique with more training data, we have performed experiments with a subset of this larger dataset. In order to ensure that the experiments are comparable with the placing task protocol, all photos from the users that appear in the test set have been removed to create the subset. The subset contains 44.8 million images, and shares 5.8 million images with the original training data (about 68% overlap). Results of experiments with this subset are

Table 4: Results showing the effect of adding additional data for enhanced modelling of geo-spatial tags and visual similarity. ME4 shows the results of an official MediaEval run.

	Percentage Precision at						Median	Error Est.
Features	100m	500m	1km	10km	100km	1000km	Error (km)	Correlation
Prior & Tags	6.17	20.52	26.55	41.84	48.47	59.22	166	0.357
Prior & LSH-SIFT	0.25	0.37	0.44	1.09	1.96	13.32	6889.5	0.152
Prior, LSH-SIFT & Tags	6.2	20.33	26.21	41.00	47.44	58.17	222.3	0.341
ME4 (Prior, LSH-SIFT & Tags)	6.66	20.55	25.97	40.6	47.04	57.97	254.5	0.372

Table 5: Results using the larger dataset without filtering photos taken by the test set users. Features key: **T**: Tags & Prior; **V**: LSH-SIFT & Prior; **H**: Tags, LSH-SIFT & Prior.

		Percentage Precision at						Median	Error Est.
	Features	100m	500m	1km	10km	100km	1000km	Error (km)	Correlation
Test set images	T	22.28	39.2	44.72	57.8	63.71	71.3	2.27	0.422
filtered	V	2.49	2.64	2.71	3.42	4.3	15.54	6759.4	0.175
	H	23.22	39.87	45.21	57.78	63.47	70.94	2.2	0.405
Test set users in	T	14.27	29.88	35.94	50.64	57.73	67.37	8.69	0.398
a 24hr window	V	0.46	0.58	0.65	1.33	2.26	13.67	6885.6	0.159
filtered	H	14.32	29.72	35.62	49.88	56.8	66.38	10.3	0.382

shown in Table 4. These results show that additional data can help, although perhaps not as substantially as might have been thought; for the tag-only run the improvement is only 3.2%. The LSH-SIFT run with the bigger data set is quite similar in performance to the baseline. Interestingly, the multimodal runs are much closer in performance to the tags-only run than they are with the baseline.

4.2.1 The effect of user modelling

Popescu [23] produced the best placing runs at MediaEval 2013 by incorporating a user-modelling step, in which all the images uploaded to Flickr by the users in the testset (with the exception of those images in a 24 hour window either side of a test image by the respective user), were crawled and used to build user models which are used in the cases where the tags don't provide high enough geographicity. We can use our model in a similar way by incorporating images from the users in the test set within our training data. Rather than crawling additional data, we have chosen to use the data we have at hand from our large Flickr dataset described above. We consider two cases: where we use all the data in the big dataset with the exception of any test images that occur (resulting in a dataset of 45.36 million images); and where we remove not only the test images, but also any other images by the same user with a time delta of less than 24 hours from the test image (dataset size of 45.12 million images). In both cases, the datasets contain 6631 users from the 7240 users in the test set; this means our data is not quite comparable to that of Popescu as he collects data for all the users. Results of experiments using both of these datasets are shown in Table 5. These results show that this additional data really helps improve performance in all cases. In terms of the tag features, the likely reason is that users often make up their own tags that are not used by other users. In terms, of the visual features, the data results indicate that users tend to take multiple images of the same scene (and at similar times) which are retrieved with very high scores by the image search system.

5. OPEN QUESTIONS AND FUTURE IDEAS

The 2013 MediaEval Placing data set provides a good baseline for researchers to compare geo-localisation techniques. One problem that is not addressed by this data set however is noise. It is well known that geotags associated with images can often be incorrect. There are many reasons for this: user error when manually labelling; errors from bulk-tagging where a user selects an entire set of photos and assigns them to a single point, when in reality they come from a region around that point; and, errors from automatic tagging, such as in the case that the GPS unit has not yet acquired a lock and thinks it is somewhere else. Dealing with these errors is a very challenging problem, however, it would be useful to create some estimate of the amount of errors in the data. One first step could be to quantify the effect of removing the *obviously incorrect* images (such as those located at exactly 0,0) from the test set and seeing by how much the techniques improve in performance.

Crandall et al. [5] did some experiments that showd that temporal features could be useful in the context of geolocalisation. It would be very easy to slot temporal features into our framework. We created some visualisations of the development data that indicated that the time-uploaded was somewhat correlated with the longitude (or perhaps timezone) in which the photo was taken. Another possibility would be to consider a temporal prior feature, which modelled a window before the query photo was uploaded/taken; there is good evidence that as time has progressed, Flickr has become more popular in Asia for example, whereas in the beginning at was only initially used in the West.

As discussed in Section 3.2, in this paper we have not discussed any form of feature weighting, other than a simple boosting of geographically related tags. This is an area ripe for future exploration: what happens if we boost tag features to the extent that visual features only have an effect for images without tags? what is the optimal weighting of the different types of feature for different placing scenarios? Popescu [23] suggested that many *machine tags* were highly location specific; what happens if we boost all machine tags in our framework?

6. CONCLUSIONS

This paper has presented a flexible approach to image geolocalisation that is based on the idea of finding the modes in a PDF. The presented technique has been shown to be

capable of directly incorporating many different types of feature. Our placing model achieves good performance without the inherent problems of models that involve splitting the Earth's surface into a grid. We have shown that adding more training data does help improve the quality of our placing models, although the improvement is relatively limited unless we include data from the user who took the query photo. We have also shown that our approach could be improved by careful selection of feature weights as demonstrated by a naïve approach to incorporating information from GeoNames.

7. ACKNOWLEDGEMENTS

The described work was funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements 270239 (ARCOMEM), and 287863 (TrendMiner).

References

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008.

[2] J. Cao. Photo set refinement and tag segmentation in georeferencing flickr photos. In [17].

[3] S. A. Chatzichristofis and Y. S. Boutalis. Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In *ICVS'08*, pages 312–322, Berlin, Heidelberg, 2008. Springer-Verlag.

[4] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *PAMI*, 24(5):603–619, 2002.

[5] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 761–770, New York, NY, USA, 2009. ACM.

[6] J. Davies, J. Hare, S. Samangooei, J. Preston, N. Jain, D. Dupplaw, and P. H. Lewis. Identifying the geographic location of an image with a multimodal probability density function. In [17].

[7] W. Dong, Z. Wang, M. Charikar, and K. Li. High-confidence near-duplicate image detection. In *ACM ICMR'12*, pages 1:1–1:8. ACM, 2012.

[8] J. S. Hare, S. Samangooei, and D. P. Dupplaw. OpenIMAJ and ImageTerrier: Java libraries and tools for scalable multimedia analysis and indexing of images. In *ACM MM'11*, pages 691–694. ACM, 2011.

[9] J. S. Hare, S. Samangooei, D. P. Dupplaw, and P. H. Lewis. Twitter's visual pulse. In *ICMR'13*, pages 297–298, New York, NY, USA, 2013. ACM.

[10] C. Hauff and G. Houben. Geo-location estimation of flickr images: Social web based enrichment. In *ECIR 2012*, pages p. 85–96. Springer LNCS 7224, April 1-5 2012.

[11] C. Hauff, B. Thomee, and M. Trevisiol. Working Notes for the Placing Task at MediaEval 2013. In [17].

[12] J. Hays and A. A. Efros. im2gps: estimating geographic information from a single image. In *CVPR' 08*, 2008.

[13] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(1):117–128, Jan. 2011.

[14] P. Kelm, S. Schmiedeke, and T. Sikora. A hierarchical, multi-modal approach for placing videos on the map using millions of flickr photographs. In *Proceedings of the 2011 ACM Workshop on Social and Behavioural Networked Media Access*, SBNMA '11, pages 15–20, New York, NY, USA, 2011. ACM.

[15] G. Kordopatis-Zilos, S. Papadopoulos, E. Spyromitros-Xioufis, A. L. Symeonidis, and Y. Kompatsiaris. Certh at mediaeval placing task 2013. In [17].

[16] O. V. Laere, S. Schockaert, and B. Dhoedt. Combining multi-resolution evidence for georeferencing flickr images. In A. Deshpande and A. Hunter, editors, *SUM*, volume 6379 of *LNCS*, pages 347–360. Springer, 2010.

[17] M. Larson, X. Anguera, T. Reuter, G. J. Jones, B. Ionescu, M. Schedl, T. Piatrik, C. Hauff, and M. Soleymani, editors. *Working Notes Proceedings of the MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013. CEUR-WS.org.

[18] L. T. Li, J. Almeida, O. Penatti, R. Calumby, D. C. G. Pedronette, M. A. Gonçalves, and R. D. S. Torres. Multimodal image geocoding: The 2013 recod's approach. In [17].

[19] X. Li, M. Larson, and A. Hanjalic. Geo-visual ranking for location prediction of social images. In *ICMR'13*, pages 81–88, New York, NY, USA, 2013. ACM.

[20] X. Li, M. Riegler, M. Larson, and A. Hanjalic. Exploration of feature combination in geo-visual ranking for visual content-based location prediction. In [17].

[21] J. Luo, D. Joshi, J. Yu, and A. Gallagher. Geotagging in multimedia and computer vision—a survey. *MTAP*, 51(1):187–211, 2011.

[22] N. O'Hare and V. Murdock. Modeling locations with social media. *Information Retrieval*, 2012.

[23] A. Popescu. Cea list's participation at mediaeval 2013 placing task. In [17].

[24] P. Serdyukov, V. Murdock, and R. van Zwol. Placing flickr photos on a map. In *SIGIR'09*, pages 484–491, New York, NY, USA, 2009. ACM.

[25] R. O. Stehling, M. A. Nascimento, and A. X. Falcão. A compact and efficient image retrieval approach based on border/interior pixel classification. In *CIKM'02*, pages 102–109, New York, NY, USA, 2002. ACM.

[26] M. Trevisiol, H. Jégou, J. Delhumeau, and G. Gravier. Retrieving Geo-Location of Videos with a Divide & Conquer Hierarchical Multimodal Approach. In *ICMR'13*, Dallas, United States, Apr. 2013. ACM.

[27] O. Van Laere, S. Schockaert, and B. Dhoedt. Finding locations of flickr resources using language models and similarity search. In *ICMR 2011*, pages 48:1–48:8, New York, NY, USA, 2011. ACM.