Boosting Prediction of Geo-location for Web Images Through Integrating Multiple Knowledge Sources

Hao Kuang, Shiai Zhu, Abdulmotaleb El Saddik
Multimedia Computing Research Laboratories, University of Ottawa, Canada
hkuan041@uottawa.ca, zshiai@gmail.com, elsaddik@uottawa.ca

ABSTRACT

Estimating geographical information of a given photo is a challenging task due to the massive spread of candidate locations on the earth. With the help of freely available geo-tagged Web images, the problem can be addressed by propagating geo-coordinates (latitude and longitude) of geo-related training data, which is obtained using document retrieval techniques. The state-of-the-art approach adopts language modeling technique to estimate the probability distribution of image associated tags in a local region. Under this framework, we propose to differentiate the tags based on the knowledge explored from multiple sources. Finally, a set of geo-informative tags are identified and further emphasized during the model learning and geo-location prediction. In addition, accurate geo-coordinates are estimated by incorporating the image visual information. Experiments on a large-scale geo-tagged Flickr image dataset demonstrate the effectiveness of proposed method at different levels of evaluation granularity.

Categories and Subject Descriptors
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing

Keywords
Image geo-location, Language model, Geo-aware tags

1. INTRODUCTION

With the emergence of different sensors integrated with portable devices (e.g., smartphone and tablet), traditional media data (e.g., image and video) are enriched with some additional information (meta-data), such as geo-location generated by built-in or external GPS devices. This new feature provides an open opportunity for the location needed services, such as travel recommendation [15], image browsing by location [2] and landmark image search [10]. Besides these applications, geo-coordinates associated with images are helpful in many other ways. In [13], Liao et al. propose a geo-aware textual feature for geo-aware image classification by propagating tags from visual and geographic neighbors. City specified visual attributes are identified through analyzing geo-tagged images in [17]. In [1], geographic information is leveraged to obtain priors about a target object and facilitate the localization of the object in a coming new image.

Many tasks benefit from the dramatically growing geographic data, nevertheless, the portion of images with accurate geo-coordinates is still relatively small (5% as reported in [5]). Thus automatic prediction of image geo-location is in high demand, and has attracted widespread interest recently. While intensive research efforts have been devoted on this direction, estimating the target location from a wide spread of candidates on the earth is not trivial. With a sheer amount of user-generated images on the Web, a straightforward way is to convert the problem to searching geo-tagged reference images. The geo-coordinates of returned results are further propagated to the query images. Both textual and visual information are explored in the literature. Visual based approaches [9, 12] model each candidate location with a set of representative images. The relevance between query images and locations is estimated by employing visual matching techniques. In [3], location-discriminative image regions are discovered using latent SVM learned on the images of each location. In [16], geo-coordinates are refined using random walk on an image graph, where edge is determined by matching the visual content. While visual feature may perform well on locations with landmarks or discriminative scenes, it cannot provide sufficient clues for localizing more general images. For example, compared to Figure 1(a), where the visual content clearly indicates the image location, Figure 1(b) is easily assigned to other locations with fountain using image matching.

In contrast to visual information, user contributed text annotations usually convey strong clues of the location where

Figure 1: Example images from Flickr.
This photo was taken, such as “Washington DC”, “Capitol Hill” in Figure 1. Thus utilizing image tags is a plausible way. In [14], a probability language model is proposed, on which many successful systems [11, 9, 6] are built. In this model, each location is represented as a tag distribution estimated from the images taken around this location. Given a query image, the associated tags are used to estimate the relevance between the query and each candidate location. As some tags such as “Capitol Hill” are more informative than tags like “lake”, the tag distribution is further modified by boosting location names presented in external manually constructed sources such as GeoNames. However, geographical names may indicate different ranges of area on the earth. For example, “Capitol Hill” is more indicative than “Washington DC” for localizing images, nevertheless, they are treated equally in [14]. Besides, user generated geo-aware tags may be not included in the database. In this paper, we propose to identify and boost the geo-aware and informative tags based on knowledge learned from multiple sources. In specific, we jointly consider GeoNames database, tag distribution among candidate regions and geographic spread of tags. In addition, the importance of tags is not only utilized in learning language model, but also in the prediction process. Finally, visual information is incorporated in the fine-grained geo-coordinates estimation. Instead of fusing the results from two modalities as in [9, 4], which is risky since visual clues are much weaker and may hurt the results, we use visual feature to search similar images only within a small local region estimated by language model. In this way, the performance at fine granularity can be improved, and meanwhile the overall performance at coarse granularity will be not degraded.

2. TAG-BOOSTED LANGUAGE MODEL

Figure 2 shows the overall framework of our approach. The world map is split into \( m \times n \) grids based on the longitude/latitude coordinate. We set the interval of both longitude and latitude as one degree. A collection of geo-tagged training images is assigned to the corresponding grids based on their geo-coordinates. Then a set of geo-aware tags are identified using knowledge from multiple sources. These tags are further emphasized in modeling the tag distribution of each grid using language model. Given a query image, the most likely grid is estimated using the learned language model and geo-informative tags. Finally, the accurate geo-locations are predicted by propagating the geo-coordinates of similar training images within the selected grid.

2.1 Probability Language Model

A grid \( l \in \mathcal{L} \) is represented as a bag of tags associated with the images assigned to \( l \). Given a query image \( x \) with a tag list \( T \), the grid \( l_x \) where \( x \) is most likely taken is determined by

\[
l_x = \arg \max_{l \in \mathcal{L}} P(l|T). \tag{1}
\]

Assuming the tags in \( T \) are generated independently, we can use Bayes’ rule and get

\[
P(l|T) \propto P(l) \times \prod_{t \in T} P(t|l). \tag{2}
\]

Then the \( P(l|T) \) is estimated by representing each grid as a multinomial distribution over tags. Given a tag vocabulary \( V \), \( P(t|l) \) is computed by counting its frequency in \( l \):

\[
P(t|l) = \frac{\#(t,l) + 1}{\sum_{l \in \mathcal{L}} \#(t,l) + 1}, \tag{3}
\]

where plus-one smoothing is used to avoid zero probability when \( t \) does not appear in \( l \). Intuitively, \( P(l|T) \) can be considered as relevance score between the query image \( x \) and a candidate grid \( l \). The prior distribution of grids \( P(l) \) is estimated to be uniform and eliminated from Equation 2. By using logarithms on Equation 2, the relevance can be estimated by

\[
\log(P(l|T)) \propto \sum_{k \in T} \log(P(t|l)). \tag{4}
\]

2.2 Geo-aware Tag Mining

As discussed in section 1, the location specified tags such as city names or landmark building names can be integrated in the language model by boosting the probability \( P(t|l) \) of these tags. To identify the geo-aware tags, we define following four measurements to determine the tag informativeness.

GeoNames. Geonames database consists of location names contributed by Web users. Similar to [14], we define weights \( w^0_t = 1 \) for tags listed in the database, and \( w^0_t = 0 \) otherwise.

IDF. In addition to the knowledge source i.e. GeoNames, tag distribution among grids is also important. One well-known weighting method is inverse document frequency (IDF), which will assign higher weights to tags appeared in less grids and penalize frequent tags. Formally, it is defined as \( w^i_t = \log(N/(\#(l : t \in l))) \), where \( N \) is the total number of grids, and \( \#(\{l : t \in l\}) \) is the number of grids including \( t \).

KL-divergence. IDF is based on the statistic of absence or presence of tags in grids, nevertheless, the distribution of tags within the grids is neglected. For instance, tags appearing in same number of grids will have same IDF values. However, the tag frequently observed in a specific grid would be more representative for that grid than others. To reflect this property, we define a measurement derived from the tag frequency \( \#(t,l) \) in grids. In specific, a tag \( t \) is more informative if there is more change of grid distribution from prior \( P(L) \) to the conditional distribution \( P(L|t) \) given tag \( t \). This can be measured using KL-divergence between the prior and conditional distributions.

\[
w^k_l = KL[P(L)||P(L|t)] = \sum_{l \in \mathcal{L}} P(l) \log\left(\frac{P(l)}{P(l|t)}\right). \tag{5}
\]

where \( P(l) \) is computed as the number of tags in \( l \) divided by the number of tags in the corpus, and \( P(l|t) \) is estimated by
2.3 Accurate Geo-coordinates Prediction

Smoothing) in the detection process using Equation 4. A weighting factor. Note that the estimated tag weights are further enhanced during prediction by replacing \( \log(P(t; l)) \) in Equation 4 with \((1 + \beta_t) \times \log(P(t; l))\), where \(\beta_t\) is a weighting factor. In this way, the tag distribution within a grid is modified to emphasize the geo-informative tags. In addition, the influence of these tags is limited in the global search, it is expected to be helpful when search is performed within a local region, where much less training images are considered. In other words, if there is a training image taken at a same place with the query image, the chance of finding it using visual feature becomes high after the search area is localized. In this paper, we first compute the spatial variance \(\sigma_t\) of images with tag \(t\). Then the weight is defined as \(w_t^* = \exp(-\beta_1 \sigma_t)\), where \(\beta_1\) is a decay factor.

After normalizing the above four measurements into range [0, 1], we assign a joint weight \(w_t\) for each tag through linear fusion. Our proposed boosted language model is defined by simply replacing \#(t, l) in Equation 3 with \((1 + \beta_t) \times \#(t, l)\), where \(\beta_t\) is a weighting factor. In this way, the tag distribution within a grid is modified to emphasize the geo-informative tags. In addition, the influence of these tags is further enhanced during prediction by replacing \(\log(P(t; l))\) in Equation 4 with \((1 + \beta_t) \times \log(P(t; l))\), where \(\beta_t\) is also a weighting factor. Note that the estimated tag weights are assigned to both present tags and absent tags (with plus-one smoothing) in the detection process using Equation 4.

2.3 Accurate Geo-coordinates Prediction

With language models, the most likely grid can be estimated according to Equation 1. Since each grid covers several hundred kilometers on the map, an accurate geo-coordinates is necessary. A straightforward way is to propagate the geo-coordinates of similar images within the selected grid to the query image. While the effectiveness of visual feature is limited in the global search, it is expected to be helpful when search is performed within a local region, where much less training images are considered. In other words, if there is a training image taken at a same place with the query image, the chance of finding it using visual feature becomes high after the search area is localized. In this paper, we use Jaccard coefficient to compute the textual similarity between two tag lists, and Cosine similarity on the visual features. The most similar image is identified using the linearly fused textual and visual similarities.

3. EXPERIMENTS

3.1 Settings and Evaluation Criteria

Our experiments are conducted on the dataset used in MediaEval placing task 2013 [7], which includes more than 8.5 million images with geo-coordinates, and several sets of testing data. After stop-words removal and stemming, we get 6.0 million images with tags. These images are assigned to the split grids according to their geo-coordinates, and finally we have 8,097 non-empty grids. Visual feature used in the accurate geo-coordinates prediction is 144 dimensional CEDD provided by MediaEval, which performs well in this task. The experiments are conducted on the first testing set including 5,300 images. To evaluate the approaches, we adopt the two standard criteria used in MediaEval. The first one is hitting rate or accuracy of predictions within a given circle radius (10 M, 100 M, 1 KM, 10 KM and 100 KM). The second one is the absolute error distance in kilometers between estimated and ground-truth coordinates.

3.2 Results and Discussions

3.2.1 Effectiveness of Identified Tags

The baseline method is the primary language model (LM) proposed in [14]. Our proposed tag-boosted approach is LM-ALL, where four measurements are jointly considered. In addition, LM-Geo, LM-IDF, LM-KL and LM-VAR are four simplified boosted language models using one of the four measurements respectively. Note that visual information is not used in this experiment. The prediction results at several evaluation radiuses are showed in Figure 3. By incorporating estimated importance of user tags, all the five boosted language modeling approaches improve the baseline significantly. Meanwhile, the improvements are consistently observed for different evaluation radiuses. LM-ALL using all the information achieves best results with 22% improvement of accuracy (radius = 1 KM) over baseline. Among the four measurements, LM-VAR using spatial variance of tags performs better than others. This is because that statistic based on geographical spread of tags is the most informative clue for localizing the query tags. The two methods LM-IDF and LM-KL using tag distribution among grids generate similar improvements over baseline, which indicates that informative tags are helpful to distinguish candidate grids. Although LM-Geo performs worse than other measurements as equal weights are assigned to the location names, it still improves the accuracy at 1 KM radius by 7% over LM. Another reason may be that statistics from user generated tags can be better to capture user tagging behaviors. Similar conclusions can be made from the average distance errors
of estimation showed in Table 1. LM-ALL significantly reduces the error from 506 KM of baseline to 129 KM. Table 2 further shows some identified tags using presented four measurements respectively. For GeoNames, we randomly select some city names from the database. Since all the training images are grouped according to their geo-coordinates, informative tags identified by KL and IDF include many location names but with different weights. In addition, spatial variance tends to select the tags (e.g., “baselstadt”) used only in a local area. These tags appearing in a query image strongly support its geo-location. Thus LM-VAR performs better than other three measurements. Best result is achieved by LM-ALL which combines the tags representing different location granularities.

### 3.2.2 Visual Feature Evaluation

To evaluate the usefulness of visual feature, we first show the result of 1NN visual matching (V-1NN) approach which is purely based on visual information of query images. As showed in Table 3, as to be expected, the estimation is extremely imprecise. This is consistent to the observations in [4, 9]. The reason is that both training and testing images are all general images rather than landmarks with strong clues indicating the location information. In addition, different from other retrieval problems, a small difference between two image matching results may result in large errors of geo-location estimation. In our method, the visual feature is leveraged in accurate geo-coordinates estimation within a local area selected by language models. We show the results of LM-V and LM-ALL-V, which respectively utilize LM and LM-ALL methods for localizing the most likely grid. For LM-V and LM-ALL-V, accurate geo-coordinate is predicted using both textual and visual features described in section 2.3. From Table 3, we can see that LM-V and LM-ALL-V incorporating visual feature improve LM and LM-ALL respectively when evaluation radius is 10 M or 100 M. Since the LM-V and LM-ALL-V are based on the initial grid localization results, the improvements at coarse evaluation granularities (larger radius) are marginal. In other words, utilizing visual feature at localized regions can improve the fine-grained prediction results without hurting the overall performance.

### 4. CONCLUSIONS

We have presented a tag-boosted language model for image geo-location prediction based on a set of geo-aware tags, which are identified from GeoNames database, statistics of tag distribution among geo-grids and the tag spatial variance. Incorporating the importance of these tags into model learning and prediction, the performance is improved significantly at different levels of evaluation granularity. In addition, visual information has been proved to be helpful in fine-grained location estimation. As image matching is helpful to the global search of certain kinds of images (e.g., landmark), we will leverage the visual information in a query-adaptive way in the future work.

### 5. REFERENCES


