Spatial-aware Multimodal Location Estimation for Social Images

Jiewei Cao[1], Zi Huang[1], and Yang Yang[2]
[1]The University of Queensland, Australia
[2]University of Electronic Science and Technology of China, China
j.cao3@uq.edu.au, huang@itee.uq.edu.au, dlyyang@gmail.com

ABSTRACT

Nowadays the locations of social images play an important role in geographic knowledge discovery. However, most social images still lack the location information, driving location estimation for social images to have recently become an active research topic. With the rapid growth of social images, new challenges have been posed: 1) data quality of social images is an issue because they are often associated with noises and error-prone user-generated content, such as junk comments and misspelled words; and 2) data sparsity exists in social images despite the large volume, since most of them are unevenly distributed around the world and their contextual information is often missing or incomplete. In this paper, we propose a spatial-aware multimodal location estimation (SMLE) framework to tackle the above challenges. Specifically, a spatial-aware language model (SLM) is proposed to detect the high quality location-indicative tags from large datasets. We also design a spatial-aware topic model, namely spatial-aware regularized latent semantic indexing (SRLSI), to discover geographic topics and alleviate the data sparseness problem existing in language modeling. Taking multi-modalities of social images into consideration, we employ the learning to rank approach to fuse multiple evidences derived from textual features represented by SLM and SRLSI, and visual features represented by bag-of-visual-words (BoVW). Importantly, an ad hoc method is introduced to construct the training dataset with spatial-aware relevance labels for learning to rank training. Finally, given a query image, its location is estimated as the location of its most relevant image returned from the learning to rank model. The proposed framework is evaluated on a public benchmark provided by MediaEval 2013 Placing Task, which contains more than 8.5 million images crawled from Flickr. Extensive experiments on this dataset demonstrate the superior performance of the proposed methods over the state-of-the-art approaches.

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806249.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models; Information filtering

Keywords

Location Estimation; Geotagging; Multimodal

1. INTRODUCTION

In recent years, driven by the proliferation of GPS-enabled devices and media-sharing platforms such as Flickr, Twitter, YouTube and Foursquare, people have been creating large collections of online multimedia content with coordinates. It has been reported that 62% of on-the-go consumers publish social sharing images, videos or posts with geo-tags, which are either manually labeled or automatically assigned by GPS-enabled devices [21]. Such georeferenced community-contributed content has been attracting extensive attention from both industry and research communities [18].

However, publicly-visible location annotations are remarkably sparse in online data. It has been estimated that only about 5% of the existing multimedia content on the Internet is actually geotagged [7]. Specifically, only 2.5% of the most-viewed videos on YouTube and approximately 4.3% of Flickr images are tagged with coordinates. Therefore, multimedia location estimation has become an active research topic. This task is also known as *geocoding* in the geographic information retrieval community and *geotagging* or *georeferencing* in the multimedia field [18]. In this paper, we are focused on the social image location estimation by utilising both the visual and textual information.

Location estimation based on visual content has been widely studied. The intuition is that the locations of two images might be close to each other if they share highly similar visual content. Early visual-based methods make the estimation within a geographically constrained area, such as urban environments or city-scale areas [29, 23]. [11] is among the first to address the images location estimation at the global level and [16] introduces a geo-visual ranking method to further improve the accuracy. Both work employ content-based image retrieval (CBIR) techniques to search the visually-nearest-neighbors with respect to a geotagged image corpus, and estimate the location of an image according to the coordinators of its neighbors.

The main drawback of the visual-based methods lies in the low prediction accuracy. They require images to contain landmarks with highly discriminative visual patterns, which rarely happens for many social images. Therefore, textual

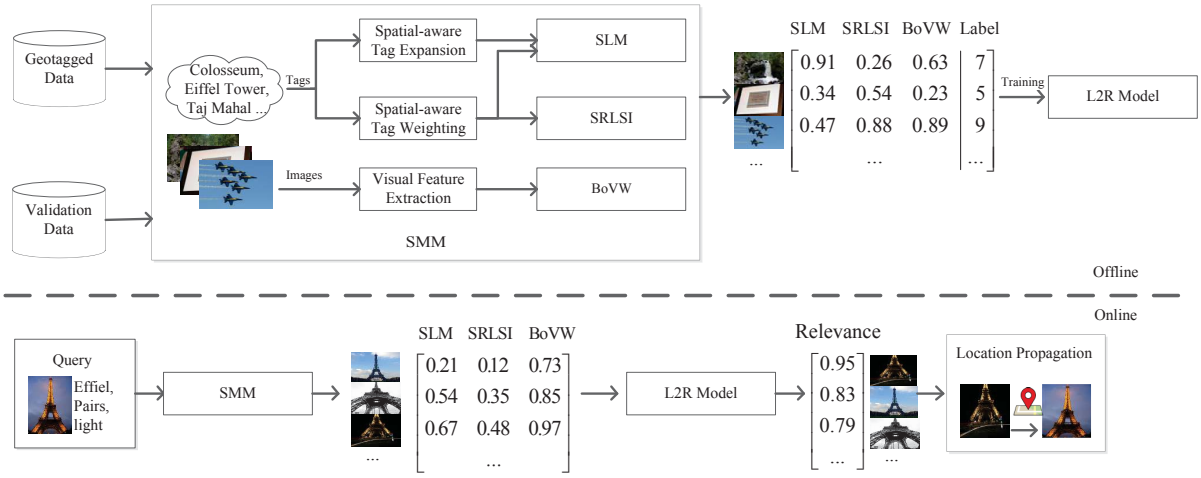


Figure 1: The proposed spatial-aware multimodal location estimation framework.

information is involved naturally, which aims to make an accurate estimation by exploiting the metadata including tags, titles and descriptions associated with social images. It has been found that language modeling (LM) approaches are particularly suitable [24, 25]. While gazetteers have traditionally been the main tool to assess the geographic scope of textual resources, they are limited to manually compiled lists of toponyms. As a compensation, large collections of georeferenced text, such as the associated tags of geotagged Flickr images, can be used to identify correlations between occurrences of terms and particular places. A vital step for these approaches is to extract spatial-aware terms [19] that are most indicative of geographic locations from a given text collection. Language model *smoothing* is required due to data sparseness. However, traditional smoothing techniques are not well tailored to the location estimation problem. In this paper, we propose a spatial-aware topic model to discover compact and readable geographic topics from georeferenced text. Integrated with language modeling, this topic model can serve as an effective smoothing technique to alleviate the data sparseness problem.

Inspired by the aforementioned observations and analysis, here we propose a spatial-aware multimodal location estimation (SMLE) framework for location estimation of social images. As illustrated in Fig.1, the framework has an offline process and an online process. In the offline process, learning to rank (L2R) is employed to combine three lists of relevance scores computed from the spatial-aware multi-modeling module (SMM) which consists of three models, namely spatial-aware language model (SLM), spatial-aware regularized latent semantic indexing model (SRLSI) and bag-of-visual-words image retrieval model (BoVW). Given a geotagged dataset, such as Flickr images and their tags, SMM handles these data by three models accordingly. Tags are preprocessed with spatial-aware tag weighting and tag expansion, and then consumed by SLM and SRLSI. And images are used by the BoVW model for visual word dictionary construction and indexing. To effectively train learning to rank, we introduce an ad hoc method to construct the training dataset with spatial-aware relevance labels. In the online process, given a query (or test) image, it is firstly processed by SMM to generate three lists of relevance scores from database images, which are then fused by the learning to rank model to generate the final ranking list. The final location estimation is conducted by propagating the top one image's location to the query image.

Our main contributions are summarized as follows: 1) the location is estimated by considering multimodal relevance scores obtained from three different models including SLM, SRLSI and BoVW; 2) the visibility of location-indicative tags is enhanced by spatial-aware tag weighting and tag expansion; 3) a novel topic model is designed to learn geographic topics from social images according to their location-indicative tags; 4) an ad hoc method is developed to construct the training dataset with spatial-aware relevance labels for learning to rank training; 5) an extensive performance study is conducted on a large-scale benchmark dataset to prove the effectiveness of our framework, in comparison with the state-of-the-art methods.

The rest of the paper is organized as follows. Related work is reviewed in Section 2. The proposed SMM and its methodologies are presented in Section 3, followed by learning to rank in Section 4. Section 5 reports the results and the paper is concluded in Section 6.

2. RELATED WORK

In this paper, we focus on the location estimation for social images on a global scale. More comprehensive surveys about geotagging in multimedia and computer vision domains and its applications can be found in [18] and [4]. In this section, we divide related work into three categories according to the modalities they utilized.

The first group is visual-based location estimation. Early work simplifies the problem to do estimation within a highly geographically constrained area. Given an urban images dataset, Zhang et al. [29] first select the best visually similar candidates based on SIFT keypoints matching and then estimate the location by performing position triangulation on top two best reference views selected by the camera motion estimation. Schindler et al. [23] work on a city-scale images dataset, and point out that the accuracy can be greatly improved by choosing the most informative visual features during vocabulary construction and efficiently increasing the branching factor in vocabulary tree searching.

Hays and Efros [11] address the image location estimation in a global level. They first collect 6 million of geotagged Flickr images and extract various visual features. For each query image, they calculate its aggregated feature distances against the whole referenced dataset in order to find the nearest-neighbors, and estimate the location according to these neighbors' coordinates. Li et al. [16] further refine the visually similar candidates by considering their geo-visual neighbors which are both geographically nearby and visually similar.

The second group is textual-based location estimation. Serdyukov et al. [24] estimate the locations of Flickr images by using a language model purely based on the tags assigned by users. They further improve the model by location-aware form of smoothing which utilized the spatial distribution of tags. Van Laere et al. [25] propose a two-step approach which first determines the most possible area for the query image based on language model and then performs similarity search within this area to find the most textually similar candidate for final location estimation. They further improve the classification accuracy by spatial-aware terms selection [19] which reduces the negative impact of noisy tags by detecting the most location-indicative tags from the whole corpus. Another approach to improve the language model is to combine it with geographic topic model. For example, Eisenstein et al. [6] propose a multi-level generative model to discover coherent topics and their regional variants which reveal topic-specific regional distinctions.

The third group is multimodal location estimation. Friedland et al. [7] introduce the concept of multimodal location estimation which leverages cues from the visual and the acoustic content of a video as well as from the given metadata. They attempt to classify which city a video comes from by matching videos containing audio from ambulance sirens in different cities. User specific modeling approaches utilize social relationship for location estimation. For example, since most social images are home-style, [13] incorporates user home location prior to favor location which is closed to user's home and shows great accuracy improvement. On the other hand, most users nowadays use Twitter to share their experiences while traveling, Hauff et al. [9] show that extracting location information from users' microblog stream can also benefit images locating.

3. SPATIAL-AWARE MULTI-MODELING

In this section, we give a detailed description of the proposed spatial-aware multi-modeling module (SMM) in the proposed spatial-aware multimodal location estimation framework, including its three models: spatial-aware language model, spatial-aware topic model, and BoVW image retrieval model. The relevance scores computed from the above three models will be used in learning to rank for both offline training and online location estimation.

3.1 Spatial-aware Language Model

To estimate the location of a query image, a straightforward solution is to find the most similar image from the database and assign its location to the query image. Given that each image is associated with a set of tags, it can be represented by its tags as a *document*. Usually, this document is short as images are generally assigned with a limited number of tags. Traditionally, this image textual representation only treats the associated tags with equal weights, which limits the advantage of taking into account the significant location information carried by the location-indicative tags. For the problem of location estimation, we improve the image textual representation by considering: 1) spatial-aware tag weighting, which assigns high weights to the location-indicative tags; and 2) spatial-aware tag expansion, which alleviate the problem that short text documents yield little in the way of term frequency information.

3.1.1 Spatial-aware Tag Weighting

In order to increase the weights of the tags that carry geographical cues, we apply the method proposed in [19] to calculate the spatial-aware weight for each tag. Given a tag t and the image set $\mathbf{L}_t$ consisting of the images with this tag, the weight of t, denoted as $s(t)$, is computed as:

$$s(t) = logN_t \cdot \frac{\sum_{\mathbf{p}\in\mathbf{L}_t}(|\{\mathbf{q}|\mathbf{q}\in\mathbf{L}_t, \mathbf{q}\neq\mathbf{p}, dist(\mathbf{p},\mathbf{q})\leq\lambda\}|)^w}{N_t^2}, \quad (1)$$

where $N_t = |\mathbf{L}_t|$ is the total number of occurrences of tag t, $\mathbf{p}$ and $\mathbf{q}$ are images in $\mathbf{L}_t$, $dist(\cdot)$ is the geographic distance (e.g., Haversine distance), λ is the distance threshold, and w is the hyper-parameter controlling the weight to favor tags whose occurrences are centered around only a few locations. The weight $s(t)$ is calculated in the way similar to "tf-idf", where the first part $logN_t$ prefers tags with high frequencies in the corpus and the second part will downgrade the $s(t)$ if tag t spreads all over the world and vise verse. Specifically, when $w = 1$, if all the images with tag t are clustered in a small region (controlled by λ), then the second part will be close to 1, otherwise, close to 0. For each tag in the corpus, we calculate its spatial-aware weight by Eq.1 and apply it in the language model and topic model later on.

3.1.2 Spatial-aware Tag Expansion

Given that each image is represented as a short document consisting of its tags, it is assumed that document closeness may indicate geographical closeness of images, if those tags are location-indicative. However, similarity search in short documents is challenging for traditional retrieval models because most tags occur only once in a single document. It is difficult to accurately estimate the relevance of documents in terms of term frequency. There are two ways to address such problem: 1) apply document clustering during language model smoothing; and 2) document expansion, which modifies each document by adding additional terms from its neighborhood documents. The latter one is particularly useful to the retrieval of short and noisy documents, where the additional terms can enrich the document content.

We propose the following method to expand the tags of each image in the whole corpus to enhance their location visibility. For each image represented by $\mathbf{d}_i$ in the image corpus $\mathbf{D}$, we find a set of its neighbors $\mathbf{D}_i$ which satisfies:

$$\mathbf{D}_i = \{\mathbf{d}_j|\forall\mathbf{d}_j\in\mathbf{D}, j\neq i, dist(\mathbf{d}_i,\mathbf{d}_j)\leqslant\lambda_1, |\mathbf{d}_i\cap\mathbf{d}_j|\geqslant\lambda_2\}, \quad (2)$$

where $dist(\cdot)$ is the geographic distance between the locations of the corresponding images $\mathbf{d}_i$ and $\mathbf{d}_j$, and $|\mathbf{d}_i\cap\mathbf{d}_j|$ is the number of their overlapping tags. We expand $\mathbf{d}_i$ by duplicating the tags which also appear in its neighborhood images. In details, for any $\mathbf{d}_j\in\mathbf{D}_i$, its overlapping tags with $\mathbf{d}_i$ will be duplicated in $\mathbf{d}_i$. For example, given $\mathbf{d}_i$={Paris, Eiffel, flower} and its two neighbors $\mathbf{d}_j^{'}$={Eiffel, night} and $\mathbf{d}_j^{''}$={Eiffel, sunny}, $\mathbf{d}_i$ will be expanded to {Paris, Eiffel,

Eiffel, Eiffel, flower} as "Eiffel" appear twice in two neighboring images. In this case, the tags in $\mathbf{d}_i$ are not unique any more. The duplicate tags in $\mathbf{d}_i$ are assigned with the same weight as computed in Eq.1 and assumed to be of high importance to their belonging images in location estimation. It can be understood that such tag expansion actually increases the frequency of location-indicative tags in image's textual representation. We conduct the spatial-aware tag expansion on all the images in the corpus $\mathbf{D}$.

3.1.3 Language Model Relevance Calculation

Given a query image $\mathbf{q}$ represented by its tags, we apply the query likelihood retrieval model [20] to calculate the relevance between the query $\mathbf{q}$ and any image $\mathbf{d}$ from the corpus where spatial-aware tag weighting and tag expansion is applied. In the query likelihood retrieval model, the language model score, denoted as $P(\mathbf{q}|\mathbf{d})$, is formally defined as follows:

$$P(\mathbf{q}|\mathbf{d}) = \prod_{t \in \mathbf{q}} (P(t|\mathbf{d}) \times s(t)), \tag{3}$$

$$P(t|\mathbf{d}) = \frac{tf_{t,\mathbf{d}} + \mu P(t|\mathbf{D})}{|\mathbf{d}| + \mu}, \tag{4}$$

where t stands for a tag, and μ is the Dirichlet prior parameter that needs to be tuned.

The language model score $P(\mathbf{q}|\mathbf{d})$ is used to measure the relevance between $\mathbf{q}$ and $\mathbf{d}$, which is based on the tag matching between the query image and the corpus images. For easy illustration in the rest of the paper, we denote the the relevance score between $\mathbf{q}$ and $\mathbf{d}$ computed from our spatial-aware language model as $s_{lm}(\mathbf{q}, \mathbf{d})$. In other words, $s_{lm}(\mathbf{q}, \mathbf{d}) = P(\mathbf{q}|\mathbf{d})$. This spatial-aware language model (SLM) is one of the three models in the proposed SMM. The relevance score computed from SLM is one of the three scores to be fused in learning to rank.

3.2 Spatial-aware Topic Model

Besides the language model score, we define the topic model score in this subsection. Essentially, the SLM is based on the tag matching (or term matching). The advantage of incorporating the topic model is to alleviate the "term mismatch" problem. Traditional text retrieval models, such as vector space model, language model and BM25, are all based on term matching. The term mismatch problem arises when the documents and the query use different terms to describe the same concept, and thus relevant documents may get low ranking scores. It is beneficial to integrate the topic matching scores with the term matching scores, to leverage both broad topic relevance and specific term relevance. Here we consider using the topic model to further improve the accuracy of textual-based location estimation. However, existing topic models are not well suited for location estimation, since they do not consider the location property of tags associated with social images.

In this subsection, we present a spatial-aware tag representation instead of using traditional term score vector (e.g., tf-idf) for latent topics learning, based on which a spatial-aware topic model named spatial-aware regularized latent semantic indexing (SRLSI) is proposed to discover geographic topics by incorporating image location factors during the topic learning phase.

3.2.1 Spatial-aware Tag Representation

Existing topic modeling methods represent a document as a M-dimensional vector, where the m^{th} entry denotes the score of the m^{th} term, such as a boolean value indicating occurrence, term frequency, tf-idf score, and joint probability of the term and document. The term score is designed to indicate the importance of a term in a specific document. Therefore, in the context of location estimation, we replace the traditional term score with the spatial-aware tag representation. Specifically, we use the tag weight calculated by Eq.1 to represent each image $\mathbf{d} = (d^{(1)}, ..., d^{(M)})$ in corpus $\mathbf{D}$ as follows:

$$d^{(m)} = \begin{cases} s(t_m) & t_m \in \mathbf{T_d} \\ 0 & \text{otherwise} \end{cases}, m \in [1, M], \tag{5}$$

where t_m is the m^{th} tag in the tag vocabulary, $\mathbf{T_d}$ is the set of tags in image $\mathbf{d}$, and M is the size of the tag vocabulary of the entire corpus. Experimental results in Section 5.3.2 show that this representation significantly improves the location estimation accuracy compared with the traditional representations.

3.2.2 Regularized Latent Semantic Indexing

There are various topic modeling methods [2], such as Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Indexing (PLSI), and Latent Dirichlet Allocation (LDA). In real-world applications, however, the usefulness of the topic modeling is often limited due to the scalability issue. Most solutions require drastic steps such as vastly reducing the size of term vocabulary to achieve efficient modeling. Regularized Latent Semantic Indexing (RLSI) [26] can scale to large datasets without reducing the term vocabulary. Given an image corpus $\mathbf{D}$, RLSI aims to solve the following optimization objective function:

$$\min_{\mathbf{U}, \{\mathbf{v}_n\}} \sum_{n=1}^{N} ||\mathbf{d}_n - \mathbf{U}\mathbf{v}_n||_2^2 + \lambda_1 \sum_{k=1}^{K} ||\mathbf{u}_k||_1 + \lambda_2 \sum_{n=1}^{N} ||\mathbf{v}_n||_2^2, \tag{6}$$

where $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_K]$ is the latent topics, K is the number of topics, $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_N]$ is the topic representations of each image in $\mathbf{D}$, and λ_1 and λ_2 are the parameters that control the effects of their regularization terms. By optimizing the above objective function, RLSI simultaneously learns the latent topics as well as topic representations. Based on RLSI, we propose the Spatial-aware RLSI (SRLSI) to calculate the geographic topic relevance between images.

3.2.3 Spatial-aware RLSI

To utilize the location cue of each image, the intuition of SRLSI is that: given two images $\mathbf{d}_i$ and $\mathbf{d}_j$, if $|\mathbf{d}_i \cap \mathbf{d}_j| > \theta_1$ and $dist(\mathbf{d}_i, \mathbf{d}_j) < \theta_2$, then the Euclidean distance $||\mathbf{v}_i - \mathbf{v}_j||_2$ should be small. In other words, if two images are geographically closed to each other and share a certain number of common tags, their latent topic representations $\mathbf{v}_i$ and $\mathbf{v}_j$ should be similar. Therefore, we propose the objective function of SRLSI as follows:

$$\min_{\mathbf{U}, \{\mathbf{v}_n\}} (\sum_{n=1}^{N} ||\mathbf{d}_n - \mathbf{U}\mathbf{v}_n||_2^2 + \lambda_1 \sum_{k=1}^{K} ||\mathbf{u}_k||_1 + \lambda_2 \sum_{n=1}^{N} ||\mathbf{v}_n||_2^2 + \lambda_3 \sum_{i,j=1}^{N} w_{ij} ||\mathbf{v}_i - \mathbf{v}_j||_2^2), \tag{7}$$

where the non-negative weight $w_{ij} > 0$ controls the effects of the latent topic representation distances among different image pairs, and λ_3 controls the effects of the whole corpus' latent topic representation similarity among all the image pairs. If $w_{ij} = 0$, the distance of $\mathbf{v}_i$ and $\mathbf{v}_j$ will not be considered. If w_{ij} is large, it will penalize the image pairs with large latent topic representation distance, and as a result, the objective function will flavor smaller $||\mathbf{v}_i - \mathbf{v}_j||_2^2$ during each optimization iteration.

By embedding location closeness in the weighted adjacency matrix $W=(w_{ij})_{i,j=1,\ldots,N}$, we are able to discover geographic topics which are geographically close. By setting a proper W, we can learn similar geographic topic representations for certain image pairs. Specifically, we set:

$$w_{ij} = \begin{cases} e^{-\frac{dist(\mathbf{d}_i,\mathbf{d}_j)}{2\sigma_{dist}^2}} & \text{if } |\mathbf{d}_i \cap \mathbf{d}_j| \geqslant \theta_1 \text{ and } dist(\mathbf{d}_i,\mathbf{d}_j) \leqslant \theta_2 \\ 0 & \text{otherwise} \end{cases}, \tag{8}$$

where the parameter σ_{dist} controls the width of the spatial neighborhoods, θ_1 is the constraint of the minimum number of overlapping tags, and θ_2 is the maximum allowable geographic distance.

We regard the above topic model as a spatial-aware topic model to discover geographic topics. The optimization of Eq.7 is convex with respect to $\mathbf{U}$ when $\mathbf{V}$ is fixed and is convex with respect to $\mathbf{V}$ when $\mathbf{U}$ is fixed. However, it is not convex with respect to both of them. Following the practice in sparse coding, we optimize the objective function by alternately minimizing it with respect to tag-topic matrix $\mathbf{U}$ and topic-image matrix $\mathbf{V}$. This procedure is summarized in Algorithm 1.

Algorithm 1 Spatial-aware RLSI

Require: $\mathbf{D} \in \mathbb{R}^{M\times N}$
1: $\mathbf{V}^{(0)} \in \mathbb{R}^{K\times N} \leftarrow$ random matrix
2: **for** $t = 1 : T$ **do**
3: $\quad \mathbf{U}^{(t)} \leftarrow Update\mathbf{U}(\mathbf{D}, \mathbf{V}^{t-1})$
4: $\quad \mathbf{V}^{(t)} \leftarrow Update\mathbf{V}(\mathbf{D}, \mathbf{U}^{t})$
5: **end for**
6: **return** $\mathbf{U}^{(T)}, \mathbf{V}^{(T)}$

Algorithm 2 $Update\mathbf{U}$

Require: $\mathbf{D} \in \mathbb{R}^{M\times N}, \mathbf{V} \in \mathbb{R}^{K\times N}$
1: $\mathbf{S} \leftarrow \mathbf{V}\mathbf{V}^T$
2: $\mathbf{R} \leftarrow \mathbf{D}\mathbf{V}^T$
3: **for** $m = 1 : M$ **do**
4: $\quad \bar{\mathbf{u}}_m \leftarrow \mathbf{0}$
5: $\quad$ **repeat**
6: $\quad\quad$ **for** $k = 1 : K$ **do**
7: $\quad\quad\quad w_{mk} \leftarrow r_{mk} - \sum_{l\neq k} s_{kl}u_{ml}$
8: $\quad\quad\quad u_{mk} \leftarrow \frac{(|w_{mk}|-\frac{1}{2}\lambda_1)_+ sign(w_{mk})}{s_{kk}}$
9: $\quad\quad$ **end for**
10: $\quad$ **until** convergence
11: **end for**
12: **return** $\mathbf{U}$

Holding $\mathbf{V}$ fixed, the update of $\mathbf{U}$ is illustrated in Algorithm 2. With $\mathbf{U}$ fixed, the update of $\mathbf{V}$ amounts to the following optimization problem:

$$\min_{\{\mathbf{v}_n\}} \sum_{n=1}^{N} ||\mathbf{d}_n - \mathbf{U}\mathbf{v}_n||_2^2 + \lambda_2 \sum_{n=1}^{N} ||\mathbf{v}_n||_2^2 + \lambda_3 \sum_{i,j=1}^{N} w_{ij}||\mathbf{v}_i - \mathbf{v}_j||_2^2. \tag{9}$$

Applying the spectral clustering graph construction [27, 28] on the last weighted sum component of Eq.9, we have:

$$\min_{\mathbf{V}} ||\mathbf{D} - \mathbf{U}\mathbf{V}||_F^2 + \lambda_2||\mathbf{V}||_2^2 + \lambda_3 Tr(\mathbf{V}\mathbf{L}\mathbf{V}^T), \tag{10}$$

where $\mathbf{L}$ is the graph Laplacian matrix. There are several variants of graph Laplacians, while we use the unnormalized graph Laplacian defined as:

$$\mathbf{L} = \tilde{\mathbf{D}} - \mathbf{W}, \tag{11}$$

where $\tilde{\mathbf{D}}$ is the diagonal matrix $\tilde{\mathbf{D}}_{ii} = \sum_j w_{ij}$. Now the update of $\mathbf{V}$ amounts to solving the optimization problem Eq.10. Let its differentiation w.r.t $\mathbf{V}$ to zero, we have:

$$(\mathbf{U}^T\mathbf{U} + \lambda_2\mathbf{I})\mathbf{V} + \lambda_3\mathbf{V}L = \mathbf{U}^T\mathbf{D}. \tag{12}$$

Eq.12 is a Sylvester equation which can be solved in parallel by parallel explicitly blocked Bartels-Stewart algorithm [8].

3.2.4 *Topic Model Relevance Calculation*

Given a query image $\mathbf{q}$ and a corpus image $\mathbf{d}$, we calculate their relevance in the geographic topic space as follows. The query is represented in the geographic topic space as:

$$\mathbf{v}_q = arg\min_{\mathbf{v}} ||\mathbf{q} - \mathbf{U}\mathbf{v}||_2^2 + ||\mathbf{v}||_2^2, \tag{13}$$

where vector $\mathbf{q}$ is a weighted representation (e.g., tf-idf, spatial-aware weighting, etc.) of the query in the tag space. $\mathbf{d}$ is represented as $\mathbf{v}_d$ in the geographic topic space in the similar way. Their geographic topic relevance, denoted as $s_{tm}(\mathbf{q}, \mathbf{d})$, is then calculated as:

$$s_{tm}(\mathbf{q}, \mathbf{d}) = \frac{\langle \mathbf{v}_q, \mathbf{v}_d \rangle}{||\mathbf{v}_q||_2 \cdot ||\mathbf{v}_d||_2}, \tag{14}$$

which will be used in the final ranking step.

3.3 BoVW Relevance Calculation

The proposed two textual-based models have been discussed in Section 3.1 and 3.2. Here we also consider the visual relevance of social images in addition to the textual relevance. The typical SIFT feature and the bag-of-visual-words (BoVW) representation are applied. To construct our visual word dictionary, we extract SIFT features from images and post-process them to RootSIFT [1] in order to boost the retrieval performance. Fast approximate k-means clustering algorithm is implemented to build the visual codebook. We build the vocabulary of one million visual words for the whole image corpus. The inverted file is used to index all the corpus images to achieve efficient retrieval.

This bag-of-visual-words image retrieval model is regarded as the third model in the SMM. The visual relevance between a query image $\mathbf{q}$ and a corpus image $\mathbf{d}$ is measured by the cosine similarity of tf-idf weighting based on the BoVW representation, denoted as $s_{bovw}(\mathbf{q}, \mathbf{d})$.

4. LEARNING TO RANK

Given a query image, we can get three lists of relevance scores generated by the three models in the SMM, as described in Section 3. How to fuse these result lists and gen-

erate the final ranking list is the question to be answered in this section.

Learning to rank (L2R) is widely applied for ranking creation in modern information retrieval system, which leverages machine learning technologies to innovate more effective ranking models [14]. In this paper, we adopt the listwise L2R model [3] to fuse three result lists. In the listwise L2R model, suppose that for a query image $\mathbf{q}$, its related candidate images $\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_n$ retrieved by the SMM, and the relevance grade labels $\mathbf{y} = \{y_1, y_2, ..., y_n\}$ for each corresponding candidate are given. The vector $\mathbf{x}_i$ for each candidate image $\mathbf{d}_i$ is a triplet of the three models' relevance scores: $\mathbf{x}_i = \langle s_{lm}(\mathbf{q}, \mathbf{d}_i), s_{tm}(\mathbf{q}, \mathbf{d}_i), s_{bovw}(\mathbf{q}, \mathbf{d}_i)\rangle$. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ represent the list of all the candidates with their relevance scores. Each $(\mathbf{X}, \mathbf{y})$ pair w.r.t a given query $\mathbf{q}$ is regarded as one training instance. Given a number of m training instances $\mathbf{I} = \{(\mathbf{X}_1, \mathbf{y}_1), ..., (\mathbf{X}_m, \mathbf{y}_m)\}$, listwise L2R can be formalized as minimizing the following empirical risk function $\hat{R}$:

$$\hat{R}(F) = \frac{1}{m}\sum_{i=1}^{m} L(F(\mathbf{X}_i), \mathbf{y}_i), \tag{15}$$

where $F(\cdot)$ is the ranking function, and $L(\cdot, \cdot)$ is the loss function that evaluates the prediction accuracy of $F(\mathbf{X})$ against the corresponding grades $\mathbf{y}$. The goal of the learning task is to automatically learn a function $\hat{F}(\cdot)$ minimizing $\hat{R}$.

The first step of learning to rank for our location estimation task is to prepare a high-quality training dataset $\mathbf{I}$ with spatial-aware relevance labels $\mathbf{y}$ for each list $\mathbf{X}$ in order to learn an effective ranking model. Most commercial search engines obtain the ground truth relevance labels by mining users' click logs during searching or browsing behaviors, which is cheap to obtain. However, in location estimation, there is no such log data for mining. Although we can obtain training data by asking human annotators to label the relevance of a candidate image $\mathbf{d}$ w.r.t a query image $\mathbf{q}$, it is extremely time-consuming to generate sufficient training data for effective learning. Here we introduce an ad hoc method to construct a training dataset with spatial-aware relevance labels for listwise L2R training in the context of location estimation.

We randomly select a number of images from the training dataset as pseudo-queries. For each pseudo-query image $\mathbf{q}$, we generate the list $\mathbf{X}$ by combining the three relevance score lists of candidate images derived from the SMM. The corresponding spatial-aware relevance label for each candidate image $\mathbf{d}_i$ is calculated as follows:

$$y_i = \sum_{g=1}^{G} \alpha_g \chi_{A_g}(dist(\mathbf{q}, \mathbf{d}_i)), \tag{16}$$

$$\chi_{A_g}(dist(\mathbf{q}, \mathbf{d}_i)) = \begin{cases} 1 & \text{if } dist(\mathbf{q}, \mathbf{d}_i) \in A_g \\ 0 & \text{if } dist(\mathbf{q}, \mathbf{d}_i) \notin A_g \end{cases}, \tag{17}$$

where G is the number of different relevance grades, α_g is the defined relevance grade value, A_g is distance interval, and χ_{A_g} is the indicator function of A_g. For example, given $A_1 = [0, 10)$ as a distance interval of 0 to 10 meters and $\alpha_1 = 10$, if $dist(\mathbf{q}, \mathbf{d}_i) \in A_1$, then y_i will be assigned with the relevance grade value of $y_i = \alpha_1 = 10$. Different relevance grades can be defined to flavor different distance intervals. We go though all pseudo-query images to generate a training dataset $\mathbf{I}$ with spatial-aware relevance labels for listwise L2R model training.

When conducting the location estimation, given a query image, we first retrieve a list of the most relevant image candidates (e.g., top 100) from each of the three models in the SMM, and then merge them into one list (i.e., $\mathbf{X}$) by combining the same image candidate's multimodal scores into a triplet vector. The non-existence of a candidate image represents zero relevance score in one list. The merged list is then taken the input of the trained L2R model to predict the final relevance score of each candidate image. The final location of the query image is estimated as the location of the candidate image with the highest relevance score computed from the trained listwise L2R model.

5. EXPERIMENT

In this section, we evaluate the performance of the proposed methods on location estimation. To this end, we use the following two metrics: 1) *accuracy at x km* is the percentage of test images estimated within x km of the ground truth location. The values of x can be 1 km, 10 km, and 100 km etc.. As the estimated location is represented by latitude and longitude, the great-circle (Haversine) distance is applied in our experiments; 2) *median error distance (MER)* is the error distance that no more than half of the test images exceed. The first metric measures the accuracy at different levels of location granularity. The median error distance, on the other hand, presents the overall performance by a single value, which is convenient for overall comparisons.

In the following, we first describe the evaluation benchmark and data preprocess. As the proposed framework involves multi-modalities in the final ranking for the location estimation, we conduct comprehensive experiments to observe the effects of each model in the SMM separately, followed by the fused results from the trained listwise L2R model. Finally, we compare our whole framework with existing methods. All the experiments are conducted on a Linux server with 40 3.0GHz CPU cores and 256GB memories.

5.1 Benchmark and Data Preprocess

Placing Task has been introduced in MediaEval[1] evaluation campaign since 2010. This task requires the participants to estimate the locations (i.e., longitude and latitude) of Flickr resources (images/videos) based on their associated tags, visual features, and metadata information. *Accuracy at x km* is used in this task for the performance evaluation. In 2013, they provided a dataset with more than 8.5 million training images and 250,000 test images [10]. We use this benchmark for evaluation in the following experiments.

For textual-based location estimation, not all the images in the training dataset are useful. We carry out two standard preliminary filter steps on it: 1) images without tags are removed; 2) Flickr users are allowed to upload multiple images with the same tags or metadata at once. In order to mitigate the negative effect of bulk upload [24], for images uploaded on the same date with identical tag set and coordinates, we only keep the first uploaded image in the training dataset as their representative.

As a result, we generate a preprocessed training dataset with 4,539,384 images. In order to avoid overfitting on test dataset, for parameters tuning of the proposed spatial-ware

[1] http://www.multimediaeval.org/

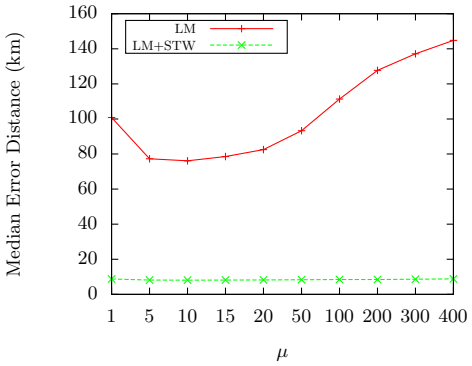


Figure 2: Comparison of the baseline (LM) and language model with spatial-aware tag weighting (LM+STW).

language model (SLM) and topic model (SRLSI), we randomly sample 50,000 images from the training dataset as a validation dataset, and the rest are used for training. Note that when separating the validation dataset from the training dataset, we ensure all the images from the same user are either in the training dataset, or in the validation dataset, in order to avoid an unfair exploitation of user-specific tags [19]. This validation dataset is also used for training the listwise L2R model. The 250,000 test images are used for the final location estimation performance evaluation.

5.2 Performance of SLM

The location of a query image is estimated as the location of the query's most relevant image from the database. To evaluate the effectiveness of the proposed SLM, we can only consider the query and database images which have textual tags. The baseline language model used for the comparison is derived by [20], where the relevance score is calculated similarly to Eq.3 by setting the weight of each tag $s(t)$ to be 1. This model treats the weights of all tags equally without the consideration of their location indication. The effects of the spatial-aware tag weighting (STW) and the performance gains after applying the spatial-aware tag expansion (STE) are discussed in the following sections respectively.

5.2.1 Spatial-aware Tag Weighting

Following [19], we set $w = 1$ and $\lambda = 40$ km in Eq.1 to favor tags that are clustered around a small number of locations. Given a query image, the baseline LM method considers all its tags as a plain text query for retrieval. In contrast, with STW, we assign different weights to each tag.

Fig.2 presents the median error distance of the baseline LM and the LM with STW regarding different Dirichlet prior μ values in Eq.4. The lowest median error distance for the baseline method is obtained when $\mu \in [5, 10]$ because most images contain five to ten tags, which means that the average text length is rather short. Comparing with the baseline LM method, we observe that LM with STW reduces the median error distance significantly and performs much stable than the baseline regarding to different μ values. This observation confirms that location-indicative tags play a much more important role than other tags in location estimation.

5.2.2 Spatial-aware Tag Expansion

We preprocess the corpus images with STE described in Section 3.1.2, and then apply the same language model (i.e., Eq.3) to find the most relevant image. The neighbor selection in Eq.2 is controlled by two thresholds: the maximum geographic distance λ_1 (in meters) and the minimum number of overlapping tags λ_2. For the ease of experiment, we discretize and restrict these two thresholds into certain values: $\lambda_1 \in [100, 200, 300, 400, 500]$ and $\lambda_2 \in [1, 2, 3, 4, 5]$.

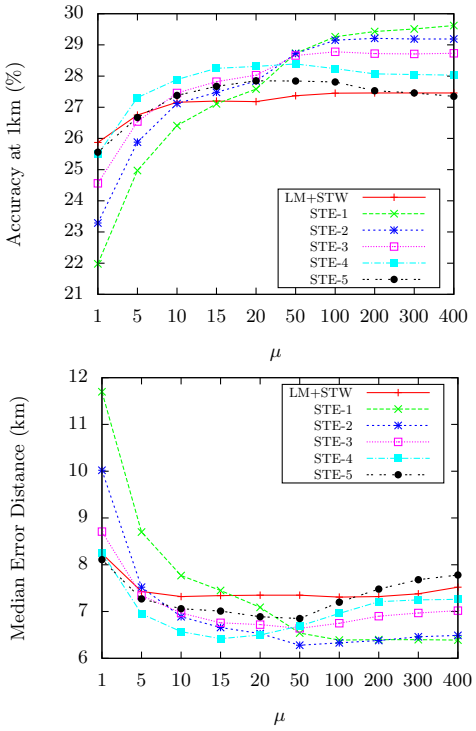


Figure 3: The influence of the parameter λ_2 in Eq.2 on spatial-aware tag expansion ("STE-k" means $\lambda_2 = k$).

Compared with the language model with STW, we study the performance improvement after applying STE. Fig.3 shows the comparisons when $\lambda_1 = 100$ for different λ_2 values (e.g., STE-1 represents the results of applying STE on LM+STW for λ_2=1). We observe similar patterns when λ_1 equals to other values and hence do not show them here. From Fig.3, as μ increase, generally speaking the accuracy increases while the median error distance drops. The reason is that STE assigns more tags to each image, which requires a larger μ for the language model smoothing. STE consistently improves LM+STW for relatively large μ values (e.g., μ >15) in terms of both accuracy and median error distance, which confirms the positive impact of increasing the frequencies of location-indicative tags in representing social images for their location estimation. However, when μ becomes very large (e.g., μ >50), the improvements for some λ_2 values start shrinking, mainly because μ gets much larger than the average number of expanded tags for each image. Considering the improvements of both accuracy and median error distance from Fig.3, we set the default values of λ_2 and μ to be 2 and 50 respectively.

5.3 Performance of SRLSI

5.3.1 SRLSI Parameters Tuning

There are three parameters controlling the weighted adjacency matrix W construction (in Eq.8): band width σ_{dist},

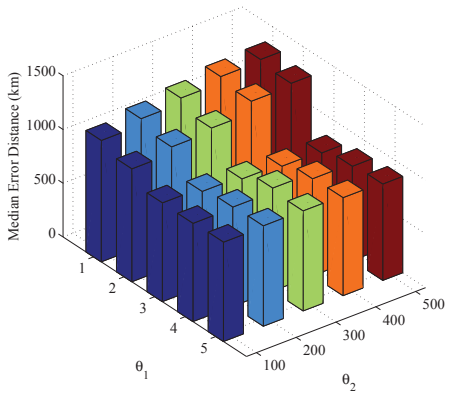


Figure 4: SRLSI performance w.r.t different θ_1 and θ_2.

minimum number of overlapping terms θ_1 and maximum geographic distance θ_2 (in meters). For the ease of experiment, we discretize and restrict these parameters into certain values: $\theta_1 \in [1, 2, 3, 4, 5]$, $\theta_2 \in [100, 200, 300, 400, 500]$ and $\sigma_{dist} \in [1, 5, 10, 15, 20]$. Similarly, we discretize the parameters λ_1, λ_2 and λ_3 in SRLSI's objective function (in Eq.7) in the range of $[0.2, 0.4, ..., 1]$. Since there are six parameters in total and we aim to improve the location estimation performance, our strategy for parameters tuning is: evaluate one or two parameters at a time and keep others fixed or assign default values, then we select the optimal values which achieve lowest median error distance on the validation dataset.

The first step is to construct matrix W. We set default values for parameters in SRLSI model: $\lambda_1 = \lambda_2 = \lambda_3 = 0.6$, and keep $\sigma_{dist} = 10$ fixed. The performance results of different combinations of θ_1 and θ_2 are shown in Fig.4. As we can see, when $\theta_1 = 1$ or 2, the median error distance is much higher than other settings. The reason is that the smaller the minimum number of overlapping tags is, the denser W is, which will introduce unexpected penalties while optimizing the objective function. The lowest median error distance is obtained when $\theta_1 = 3$ and $\theta_2 = 500$. We obtain the optimal values for the other four parameters in a similar way, and the lowest median error distance is obtain when $\sigma_{dist} = 20$, $\lambda_1 = 0.6$, $\lambda_2 = 0.4$ and $\lambda_3 = 0.4$. We use these settings in the following experiments.

5.3.2 *Spatial-aware Tag Representation*

Here, we compare SRLSI with the traditional LSI and RLSI as the baselines. For different topic models, we compare their performances with different tag representations, including tf-idf representation $\mathbf{D}_{tfidf}$ and spatial-aware tag representation $\mathbf{D}_{spatial}$. The superiority of our proposed spatial-aware tag representation outperforms the traditional tf-idf representation across all topic models, which is explicitly shown in Table 1. Besides, our SRLSI achieves the best estimation accuracy and the lowest median error distance compared with LSI and RLSI. This demonstrates the effectiveness of our spatial-aware topic model which incorporates location factors during latent geographic topic learning.

SRLSI's advantages in terms of topic readability can be observed from Table 2 which contains the topics generated by non-spatial-aware topic model LSI-$\mathbf{D}_{tfidf}$ and the topics generated by SRLSI-$\mathbf{D}_{spatial}$. We randomly show five topics with their top five weighted tags. It is obvious that the topics generated by SRLSI-$\mathbf{D}_{spatial}$ are less noisy and more relevant to specific locations. For example, the tags in Topic 1 generated by SRLSI-$\mathbf{D}_{spatial}$ are related to UK, while the tags in the topics generated by LSI-$\mathbf{D}_{tfidf}$ are less geo-related. The effectiveness of the proposed spatial-aware topic model is visualized in Fig.5. For each topic in Table 2, we randomly select 500 training images that contain its top five tags. Each image is represented by a spot on the figure. Different colors and shapes of the spots stand for different topics that the corresponding images belong to. By applying our SRLSI topic model, the images which are geographically close mostly share the same topic. It is a strong evidence to support that the SRLSI topic closeness indicates the geographical closeness of their images.

Table 1: Performance comparisons of topic models on the validation dataset, including accuracy at x km and mean error distance (ME). LSI-$\mathbf{D}_{tfidf}$ means LSI topic model with tf-idf representation, and so on.

	acc@1km	acc@10km	acc@100km	ME (km)
LSI-$\mathbf{D}_{tfidf}$	5.15	18.61	26.53	1263.93
LSI-$\mathbf{D}_{spatial}$	7.33	22.49	32.47	798.38
RLSI-$\mathbf{D}_{tfidf}$	6.30	20.40	28.13	1154.09
RLSI-$\mathbf{D}_{spatial}$	7.17	24.36	34.40	711.66
SRLSI-$\mathbf{D}_{tfidf}$	8.36	24.52	34.36	942.99
SRLSI-$\mathbf{D}_{spatial}$	**9.43**	**24.75**	**34.83**	**683.06**

Table 2: Topics learned by LSI-$\mathbf{D}_{tfidf}$ and SRLSI-$\mathbf{D}_{spatial}$.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
LSI-$\mathbf{D}_{tfidf}$	budapest	castle	croatia	hagiasophia	scotland
	hungary	nikon	dubrovnik	streetart	baltimore
	2007	2009	fountain	mannheim	glasgow
	pisa	athens	czechrepublic	badenwürttemberg	edinburgh
	lasvegas	greece	sign	ny	arizona
SRLSI-$\mathbf{D}_{spatial}$	london	edinburgh	dublin	netherlands	atlanta
	england	greenwich	ireland	holland	dragoncon
	uk	thames	salzburg	nederland	finland
	londres	england	montreal	amsterdam	helsinki
	unitedkingdom	greaterlondon	dublincastle	paysbas	montmartre

Table 3: Visual-based location estimation performance comparison on the test dataset.

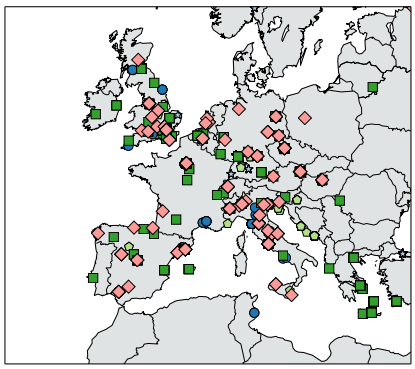
	acc@1km	acc@10km	acc@100km
SOTON [5]	0.34	0.56	0.10
RECOD [15]	0.37	0.80	1.69
CERTH [12]	0.76	1.16	2.04
Delft [17]	**2.80**	**3.70**	**4.70**
ours	2.04	2.75	3.84

5.4 Performance of BoVW

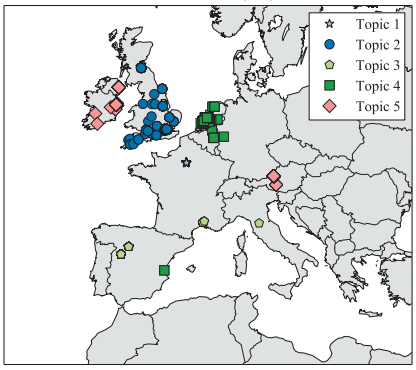
We evaluate the BoVW model directly on the test dataset without tuning any parameter. Table 3 compares the BoVW model with MediaEval 2013 Placing Task participants' best visual-based location estimation results which are also evaluated based on this benchmark. As we can see, although the BoVW model achieves comparable performance compared with the best result [17], the accuracy of visual-based location estimation is rather low, since most of the images do not have visually similar images within their geo-neighborhood as described in [16]. Therefore, visual representations can only be used as supplementary information for location estimation in this benchmark.

5.5 Performance of L2R

In this experiment, we study the performance of fusing multi-modalities using the listwise L2R model. As described in Section 4, in order to construct the training dataset for



(a) LSI-$\mathbf{D}_{tfidf}$



(b) SRLSI-$\mathbf{D}_{spatial}$

Figure 5: Location distribution of the 5 topics in Table 2. (a) is the distribution of topics generated by LSI-$\mathbf{D}_{tfidf}$; (b) is the distribution of topics generated by SRLSI-$\mathbf{D}_{spatial}$.

Table 4: Predefined relevance grades α and the corresponding distance intervals A in Eq.16.

g	1	2	...	10	11
Interval A_g (in meter)	$[0, 100]$	$(100, 200]$	...	$(900, 1000]$	$(1000, \infty)$
Grades α_g	10	9	...	1	0

the listwise L2R model, we sample 50,000 images as pseudo-queries from the whole training dataset. We define the relevance grades as shown in Table 4. For example, given a query $\mathbf{q}$, if the geographic distance between $\mathbf{q}$ and its relevant image candidate $\mathbf{d}_i$ is within 100 meters ($dist(\mathbf{q}, \mathbf{d}_i) \in A_1$), then the relevant grade value of this candidate image is $y_i = \alpha_1 = 10$ according to Eq.16. The higher grade an image has, the more relevant the image is. We select and merge the top 100 candidates retrieved by each models in the SMM, and remove candidates whose grade label is zero in order to avoid noises and reduce the size of training dataset. We eventually obtain a training dataset with 2.6 million labeled data for training the listwise L2R model.

We choose ListNet [3] as our listwise L2R model and evaluate the influence of individual model on the final location estimation by incorporating one or more models at a time during the training. Table 5 shows the performance of the trained listwise L2R model tested on the validation dataset with different model combinations. An obvious observation is that all the models and their combinations consistently reach higher estimation accuracies at larger location granularities, which is easy to understand. For two textual-based location estimation models, SLM achieves much better performance than SRLSI does, which demonstrates that the specific tag relevance matching is more capable to precisely locate an image than the topic relevance matching. The BoVW model alone has the worst estimation performance as expected. From the above observations, we consider topic relevance and visual relevance as supplementary information when performing models fusion in listwise L2R model. When SLM combines with SRLSI or BoVW, the combination achieves a higher accuracy in locating images within 1 km and a lower median error distance. When all the models are combined, we achieve the highest estimation accuracy and the lowest median error distance. This means we can locate the images more precisely by incorporating more information from other models. Especially, the improvement on smaller location granularities (e.g., 1 km) are more significant. In practice, location estimation within a finer location granularity is more useful for real-world applications.

Table 5: Multimodal performance comparisons on the validation dataset. *SLM*, *SRLSI*, and *BoVW* are each model's performance before fusion. *SLM+SRLSI* means both SLM score list and SRLSI score list are used in the listwise L2R model, etc. *All* means all three models' score lists are used in the listwise L2R model.

	acc@1km	acc@10km	acc@100km	ME(km)
SLM	28.72	53.63	66.29	6.28
SRLSI	11.21	25.41	38.67	338.72
BoVW	1.99	2.55	3.42	5391.87
SLM + SRLSI	29.31	53.52	65.21	6.23
SLM + BoVW	29.13	53.69	65.95	6.24
All	**31.22**	**54.13**	**67.05**	**5.89**

5.6 Comparison with Existing Methods

In this experiment, we evaluate our proposed framework SMLE's location estimation performance on the whole test dataset containing 250,000 images, compared with [13] and the methods developed by all the groups participating in MediaEval 2013 Placing Task. For a fair comparison, we report the best results achieved by each method that are based on the provided 8.5M training dataset only, without using external data.

Table 6 shows the comparison results on the whole test dataset. The CERTH [12] team uses both visual and textual data, while the others use textual data only. Our proposed framework SMLE surpasses all the methods in terms of both the estimation accuracy and the median error distance. When compared with the CEA LIST's method and the GHENT's method, which also apply the language model to find probable location, our spatial-aware multimodal framework improves the estimation accuracy by up to 34% relatively, and reduces the median error distance significantly from the second best performer GHENT's 51.07 to 32.29km only. This demonstrates the benefits of incorporating more information from different model. The CERTH's method simply applies a fallback strategy to combine the visual and textual information: if the test image had no tag associated with it, then the visual feature is employed to find the most relevant neighbor. In contrast, our multimodal relevance scores are fused and candidate images are ranked by the trained listwise L2R model, which provides much better estimation results. The above comparisons verify the supe-

Table 6: Performance comparison on the test dataset.

	acc@1km	acc@10km	acc@100km	ME(km)
CERTH [12]	10.37	23.70	36.22	681.00
SOTON [5]	23.15	37.70	43.82	451.89
CEA-LIST [22]	26.00	43.00	50.00	98.80
GHENT [13]	20.26	41.96	53.15	51.07
SMLE	**27.22**	**44.99**	**54.98**	**32.29**

riority of the proposed spatial-aware multimodal framework in terms of both the estimation accuracy and the median error distance. Note that the performance of our framework on the test dataset in Table 6 is not as good as the performance when tested on the validation dataset in Table 5. The reason is that all the images in the validation dataset are associated with tags while 13% images in the test dataset have no tags at all, which suggest that we can only estimate those non-tagged images via visual relevance ranking only.

6. CONCLUSION

In this paper, we propose a spatial-aware multimodal location estimation framework to address the social image's location estimation problem. We propose a spatial-aware language model to improve textual-based location estimation by utilizing spatial-aware tag weighting and tag expansion. We also design a spatial-aware topic model to discover geographic topics which can further facilitate the estimation. A learning to rank model is adopted and effectively trained to fuse multiple textual and visual models to generate a refined ranking list for final location estimation. Extensive experiments illustrate the advantages of the proposed framework over the state-of-the-art approaches.

7. REFERENCES

[1] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, pages 2911–2918, 2012.

[2] D. M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, 2012.

[3] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *ICML*, pages 129–136, 2007.

[4] J. Choi and G. Friedland. *Multimodal Location Estimation of Videos and Images*. Springer, 2015.

[5] 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 *MediaEval*, 2013.

[6] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *EMNLP*, pages 1277–1287, 2010.

[7] G. Friedland, O. Vinyals, and T. Darrell. Multimodal location estimation. In *ACM MM*, pages 1245–1252, 2010.

[8] R. Granat and B. Kågström. Evaluating parallel algorithms for solving sylvester-type matrix equations: Direct transformation-based versus iterative matrix-sign-function-based methods. In *PARA*, pages 719–729, 2004.

[9] C. Hauff and G.-J. Houben. Placing images on the world map: a microblog-based enrichment approach. In *SIGIR*, pages 691–700, 2012.

[10] C. Hauff, B. Thomee, and M. Trevisiol. Working notes for the placing task at mediaeval 2013. In *MediaEval*, 2013.

[11] J. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *CVPR*, pages 1–8, 2008.

[12] G. Kordopatis-Zilos, S. Papadopoulos, E. S. Xioufis, A. L. Symeonidis, and Y. Kompatsiaris. CERTH at mediaeval placing task 2013. In *MediaEval*, 2013.

[13] O. V. Laere, S. Schockaert, and B. Dhoedt. Georeferencing flickr resources based on textual meta-data. *IS*, 238(0):52 – 74, 2013.

[14] H. Li. Learning to rank for information retrieval and natural language processing. *Synthesis HLT*, 4(1):1–113, 2011.

[15] L. T. Li, J. Almeida, O. A. B. Penatti, R. T. Calumby, D. C. G. Pedronette, M. A. Gonçalves, and R. da Silva Torres. Multimodal image geocoding: The 2013 RECOD's approach. In *MediaEval*, 2013.

[16] X. Li, M. Larson, and A. Hanjalic. Geo-visual ranking for location prediction of social images. In *ICMR*, pages 81–88, 2013.

[17] X. Li, M. Riegler, M. Larson, and A. Hanjalic. Exploration of feature combination in geo-visual ranking for visual content-based location prediction. In *MediaEval*, 2013.

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

[19] O. Van Laere, J. A. Quinn, S. Schockaert, and B. Dhoedt. Spatially aware term selection for geotagging. *TKDE*, 26(1):221–234, 2014.

[20] D. Metzler and W. B. Croft. Combining the language model and inference network approaches to retrieval. *IPM*, 40(5):735–750, 2004.

[21] NinthDecimal. Ninthdecimal mobile audience Q2 2012 insights report, 2012.

[22] A. Popescu. CEA LIST's participation at mediaeval 2013 placing task. In *MediaEval*, 2013.

[23] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *CVPR*, pages 1–7, 2007.

[24] P. Serdyukov, V. Murdock, and R. van Zwol. Placing flickr photos on a map. In *SIGIR*, pages 484–491, 2009.

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

[26] Q. Wang, J. Xu, H. Li, and N. Craswell. Regularized latent semantic indexing: a new approach to large-scale topic modeling. *TOIS*, 31(1):1–44, 2013.

[27] Y. Yang, Y. Yang, Z. Huang, H. T. Shen, and F. Nie. Tag localization with spatial correlations and joint group sparsity. In *CVPR*, pages 881–888, 2011.

[28] Y. Yang, Z.-J. Zha, Y. Gao, X. Zhu, and T.-S. Chua. Exploiting web images for semantic video indexing via robust sample-specific loss. *TMM*, 16(6):1677–1689, 2014.

[29] W. Zhang and J. Kosecka. Image based localization in urban environments. In *3DPVT*, pages 33–40, 2006.