Exploiting Spatial Relationship between Scenes for Hierarchical Video Geotagging

Yifang Yin, Luming Zhang, Roger Zimmermann
School of Computing, National University of Singapore, Singapore 117417
{yifang,zglumg,rogerz}@comp.nus.edu.sg

ABSTRACT
Predicting the location of a video based on its content is a very meaningful, yet very challenging problem. Most existing work has focused on developing representative visual features and then searching for visually nearest neighbors in the development set to achieve a prediction. Interestingly, the relationship between scenes has been overlooked in prior work. Two scenes that are visually different, but frequently co-occur in same location, should naturally be considered similar for the geotagging problem. To build upon the above ideas, we propose to model the geo-spatial distributions of scenes by Gaussian Mixture Models (GMMs) and measure the distribution similarity by the Jensen-Shannon divergence (JSD). Subsequently, we present the Spatial Relationship Model (SRM) for geotagging which integrates the geo-spatial relationship of scenes into a hierarchical framework. We segment the Earth's surface into multiple levels of grids and measure the likelihood of input videos with an adaptation to region granularities. We have evaluated our approach using the YFCC100M dataset in the context of the MediaEval 2014 placing task. The total set of 35,000 geotagged videos is further divided into a training set of 25,000 videos and a test set of 10,000 videos. Our experimental results demonstrate the effectiveness of our proposed framework, as our solution achieves good accuracy and outperforms existing visual approaches for video geotagging.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms
Algorithm, Experimentation

Keywords
Video geotagging, visual approach, scene distribution modeling, spatial relationship analysis

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

ICMR ’15, June 23 - 26, 2015, Shanghai, China.
Copyright 2015 ACM. 978-1-4503-3274-3/15/06$15.00.
http://dx.doi.org/10.1145/2671188.2749354.

1. INTRODUCTION
Nowadays, video recording and sharing have become popular activities in people’s lives. Due to the increasing number of sensor-equipped cameras and smartphones, geo-metadata can be automatically recorded together with multimedia content. The geo-context is of great importance, as it facilitates many location-based applications such as annotation [19, 25] and retrieval [24]. However, the number of geotagged videos is still very limited on content sharing platforms. According to Song et al. [21], only 2.5% of the most-viewed videos on YouTube are geotagged. Therefore, predicting the geographic location of where a video was captured has become an urgent yet challenging task in recent years. In the MediaEval benchmarking initiative, the Placing Task is held every year to capture the challenge of estimating the geo-locations of multimedia items. Since the geo-coordinates associated with multimedia items, which are used as the ground truth, do not always mark the precise locations, the methods are usually evaluated at different geographical margins of error of 0.1 km, 1 km, 10 km, 100 km, 1,000 km, and 10,000 km [11]. The process of estimating a capture location is often known as geotagging in the multimedia field [15].

Estimating the geo-location of a video can be achieved based on visual features [17, 14], textual information [18], social context [21], and their combinations [12, 23, 20]. Considering that some textual and social metadata used in multimodal frameworks are not always available in real life, visual approaches for video geocoding are important because they infer the geo-coordinates of a video using the visual content only. Based on the idea that videos located adjacenty share similar visual content, efforts have been made to develop compact and high-level video representations to improve geocoding accuracy [9, 17]. In the most recent geotagging approaches, spatial segmentation is usually adopted [10, 18]. This method maps the original task into a classification problem. Considering the difference in granularity between spatial cells and input videos, new challenges have been revealed in feature pooling and matching techniques.

In the most recent visual approaches for video geocoding, the Bag-of-Scenes (BoS) model has been shown to be simple and effective [17, 12]. This approach first generates a dictionary of scenes, each of which represents a specific semantic concept. Next, it assigns frames to one or more visual scenes, followed by a pooling step to generate the final representation. However, one limitation is that such methods assume the scenes in the dictionary to be independent. More emphasis should be placed on modeling the relation-
ship among scenes to achieve better performance. In this study, we analyze the geo-distributions of scenes in different spatial segments. The raw geo-coordinates of frames are replaced and smoothed with Gaussian distributions. The reason for substituting the geo-coordinates with continuous kernel functions is to reduce the sensitivity to noise in the data. Subsequently, we measure the spatial similarity of scenes based on their distributions using the Jensen-Shannon divergence (JSD), and discuss its use in geotagging problems. The contributions of this paper are twofold: (1) we model the connections between scenes by comparing their geo-spatial distributions, and (2) we integrate the spatial relationship of scenes into a hierarchical geotagging framework that can achieve better accuracies. The basic idea is that scenes that show similar geo-spatial patterns are more likely to be captured in the same video. As shown in Figure 1, our proposed geotagging system hierarchically segments the Earth’s surface into grids with different granularities. We not only pool the visual features, but also measure the spatial similarity of scenes in each of the cells in the hierarchy. Given a test video, we extract the same visual features, and estimate the posterior probability based on both the visual similarity and the spatial consistency. Using a dataset of 35,000 Flickr videos, we show that our method outperforms existing visual approaches by exploiting the spatial patterns of scenes.

![Hierarchical Framework](image)

Figure 1: The proposed hierarchical framework for video geotagging.

The rest of this paper is organized as follows. We first report on the important related work in Section 2. Next, we discuss the limitations of existing approaches and present our solution in Section 3. We introduce the distribution modeling and similarity estimation of scenes in Section 4. The proposed hierarchical framework is introduced in Section 5. The thorough experimental results of Section 6 validate the effectiveness of our system. Section 7 concludes and suggests future work.

2. RELATED WORK

Video geotagging has become a popular topic in recent years. Started from 2010, MediaEval has offered a placing task as an annual competition of web video geocoding [11]. Leveraging the textual annotations and visual features, approaches have been proposed to map the original placing task into a classification problem. Serdyukov et al. [18] segmented the Earth’s surface into an \( m \times n \) grid, and then estimated a language model from the textual annotations in each cell. After applying smoothing techniques to the spatial ambiguity, they were able to classify a test video into the cell with the highest probability. Kelm et al. [10] proposed a hierarchical spatial segmentation framework for video geotagging. Not only textual information but also visual features were pooled for each spatial segment at different hierarchy levels. Hays and Efros [7] proposed a purely visual approach that estimates image locations as a probability distribution over the globe. Penatti et al. [17] followed this line of thought and proposed an approach called Bag-of-Scenes for video location prediction. Song et al. [21] proposed to utilize the social relationship between web videos to propagate the geotags. They established this relationship graph by considering semantically related and same-author videos. Multimodel geocoding frameworks [12, 23, 6, 20] have also been proposed that process multiple sources of information for effective location estimation. Trevisio et al. [23] presented a divide & conquer strategy to better exploit the tags associated with videos. A multimodal approach was adopted which processes the information sources by decreasing order of expected informativeness: tags, user’s upload history and social graph, user’s personal information (home town) and visual content. Hare et al. [6] estimated a continuous probability density function (PDF) over the surface of the Earth from multiple features including location prior, tags, and a number of visual features. By simultaneously taking the social, visual, and textual relationships into consideration, Song et al. [20] modeled the geotagging problem as a propagation of the geography information among the web video social networks.

The task of predicting a video location by exclusively using the visual content has drawn much research attention in the computer vision community. However, according to Kelm et al. [10], visual features alone show low correlation with locations and a purely visual approach achieves limited precision. Traditionally videos can be represented by global features such as color, texture, and edge descriptors [10], local features such as SIFT and SURF [14], or motion features such as HMP [2]. To reduce the high dimensionality of visual features, methods have been developed to aggregate image descriptors into compact codes while preserving accuracy [9, 23]. As Penatti et al. [17] pointed out, scenes are elements with more semantic information in videos. Therefore, they presented a high-level video representation based on a dictionary of scenes. The distance between videos can be calculated by the Euclidean norm. Since videos located adjacently usually share similar content, most visual approaches estimate the location of a test video based on its most visually similar neighbors in the development set. We can either use the geo-coordinate of the first nearest neighbor as the prediction [7, 12, 17], or interpret it as a voting result of the top \( k \) nearest neighbors [7, 23]. For the prediction of image locations, Li et al. [14] modeled a location as an area and ranked images by comparing among their geo-visual neighbors as well. Thus, they were able to enrich the visual aspects of a location and therefore achieved better performance.

3. VIDEO LOCATION MODELING

The problem of estimating the geotag associated with a video can be seen as determining the location \( g \) with the highest probability where the video \( v \) was recorded,

\[
\arg\max_g P(g|v)
\]
By applying Bayes’ law, the posterior probability $P(g|v)$ can be formulated as Eq. 2 if we assume that the video prior $P(v)$ is uniformly distributed.

$$P(g|v) \propto P(v|g)P(v)$$  

Here $P(v|g)$ denotes the likelihood of video $v$ given location $g$, and $P(v)$ represents the prior probability of location $g$ where a video was captured.

Researchers have attempted to find a good estimation of $P(v|g)$ based on various assumptions. Next, we will briefly introduce the previous visual approaches, most of which have only considered the visual similarity between video $v$ and location $g$ when estimating $P(v|g)$. Subsequently, we will discuss existing issues and then illustrate our proposed solution.

### 3.1 Visual Similarity

Traditionally, a location has been modeled by a single photo [7]. The visual similarity between a test video $v$ and a training video with a geotag of $g$ was used as an estimate for $P(v|g)$. In image geotagging, Li et al. [14] argued that one photo only covers limited visual aspects of a location, and therefore they modeled a location as an area with the set of images in it. A similar idea has also been applied in video geotagging. Keln et al. [10] proposed to hierarchically segment the Earth’s surface and pool visual features for each spatial segment using the mean value. Thereafter, regions and videos can be directly compared and the most visually similar spatial segment is determined iteratively by calculating the Euclidean norm. Although promising results were reported, one major issue with this approach is that it ignores the granularity difference between regions and videos. This simplification hinders achieving a better accuracy.

To illustrate this problem, let us adopt the BoS [17] model for the high-level visual descriptors of videos. Given a dictionary of scenes $S = \{s_1, s_2, \ldots, s_n\}$, a video can be represented by a vector whose length equals the dictionary size, $X_v = \{x_v^1, x_v^2, \ldots, x_v^n\}$. Each element $x_v^i$ can be interpreted as the saliency score of its corresponding scene $s_i$ based on its appearance in the video. Similarly, the BoS descriptor can also be generated for a region by coding and pooling all the frames in it. However, the salient scenes that appear frequently in a region are not guaranteed to have any overlaps due to the coarse granularity. As illustrated in Figure 2, the peaks represent the distributions of two scenes in a cell. For two scenes that have no overlaps at all in geo-distributions (see Figure 2a), they can rarely be captured in a same video. Therefore, given a test video which contains both of the scenes, it is more likely taken in a cell as shown in Figure 2b. However, the existing approaches have taken little consideration of scene distributions while comparing a test video to spatial segments.

### 3.2 Spatial Consistency

In our approach we also model a location as an area. To solve the issues discussed above, we propose the Spatial Relationship Model (SRM) which combines spatial consistency with visual similarity when estimating $P(v|g)$ in Eq. 2 as follows

$$P(v|g) = SpatialCon_{v,g} \cdot VisualSim_{v,g}$$

where $VisualSim_{v,g}$ denotes the visual similarity between video $v$ and region $g$, and $SpatialCon_{v,g}$ represents the spatial consistency of the scenes captured by video $v$ in region $g$. Here we model $SpatialCon_{v,g}$ by

$$SpatialCon_{v,g} = X_v M^g X_v^\top$$

where $X_v = \{x_v^1, x_v^2, \ldots, x_v^n\}$ denotes the video’s BoS representation, and $M^g$ is an $n \times n$ matrix whose element $m^g_{ij}$ represents the geo-spatial similarity between the two scenes $s_i$ and $s_j$ in region $g$. A larger $m^g_{ij}$ means the two scenes are more likely to be captured at the same locations in region $g$. Thus, we are able to promote the spatial segments as illustrated in Figure 2b while depressing the ones as shown in Figure 2a. Next, we will first discuss how to compute the matrix $M^g$ in Section 4, and then introduce a hierarchical framework for geotagging in Section 5.

### 4. DISTRIBUTION OF SCENES

The location of a video is closely related to the visual scenes of its content. However, most of the existing work has only focused on developing representative visual features, and has searched for nearest neighbors with little consideration for the relationship between scenes. To compute the geo-spatial similarity matrix $M^g$, we first model the scene distribution using GMMs, and then estimate the pairwise geo-spatial similarity scores using JSD.

#### 4.1 Modeling Scene Distribution

We adopt the BoS [17] model as the high-level video representation. Given a dictionary of scenes, we quantize every video frame to its visually nearest scene through hard coding. Thereafter, each scene (cluster) is formed by a set of frames whose geo-coordinates are likely to be uneven distributed. As the geotags associated with frames are discrete and noisy, we smooth the raw locations in each of the clusters to create continuous distributions by the Gaussian Mixture Models as follows

$$p(x|\lambda) = \sum_{i=1}^{n} \omega_i g(x|\mu_i, \Sigma_i)$$

where $x = (lat, lon)^\top$ is a two dimensional vector representing the geo-coordinates of a frame, $\omega$ denotes the mixture weights s.t. $\sum_{i=1}^{n} \omega_i = 1$, and $g(x|\mu_i, \Sigma_i)$ represents a 2D Gaussian distribution

$$g(x|\mu_i, \Sigma_i) = \frac{1}{2\pi|\Sigma_i|^\frac{1}{2}} e^{-\frac{1}{2} (x-\mu_i)^\top \Sigma_i^{-1} (x-\mu_i)}$$

To estimate the parameters $\lambda$ of the GMMs, we derive the likelihood function as

$$L(\lambda|X) = p(X|\lambda) = \prod_{j=1}^{m} p(x_j|\lambda)$$
or the more convenient log-likelihood function as

$$ l(\lambda | X) = \log L(\lambda | X) = \sum_{j=1}^{m} \log p(x_j | \lambda) $$

(8)

where $X$ represents the geo-coordinates of the frames quantized to a cluster, and $m = |X|$ is the cardinality of set $X$. The target is to find the parameters that can best match the distribution of the training samples, that is,

$$ \underset{\lambda}{\text{argmax}} \ l(\lambda | X) $$

(9)

This problem can be solved by applying the off-the-shelf expectation maximization (EM) algorithm [5]. The EM algorithm is an iterative method: it alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found in the E step. The number of confederate Gaussian distributions $n$ can be automatically decided by recruiting a $v$-fold cross-validation [1].

By replacing the geo-coordinates with continuous kernel functions such as Gaussian probability distributions, we are able to create summary statistics that are less sensitive to high-frequency noise in the data. Such techniques have been widely used in geographical information systems [3].

4.2 Estimating Geo-spatial Similarity

In different areas, the spatial relationship of scenes is also divergent accordingly. Therefore, our next task is to calculate a good estimate of $M^g$ in each spatial segment of the Earth’s surface. As the geographic distributions of scenes have already been modeled by GMMs, the Jensen-Shannon divergence can be a good choice for estimating the geo-spatial similarity between scenes. JSD is a widely used approach to measure the similarity between two probability distributions. As shown in Eq. 10, it is a symmetrized and smoothed version of the Kullback-Leibler divergence (KLD).

$$ D_{JS} (P \parallel Q) = \frac{1}{2} D_{KL} (P \parallel M) + \frac{1}{2} D_{KL} (Q \parallel M) $$

(10)

Here $M = \frac{1}{2}(P + Q)$. For distributions $P$ and $Q$ of a continuous random variable, the KLD is defined to be the integral

$$ D_{KL} (P \parallel Q) = \int_{-\infty}^{\infty} \ln \left( \frac{p(x)}{q(x)} \right) p(x) \, dx $$

(11)

where $p(x)$ and $q(x)$ represent the densities of the distributions $P$ and $Q$. The KLD between two Gaussian mixture models can be estimated by the Monte-Carlo algorithm [8]. In our system, we utilized the Java library jMEF\(^1\) which can create and manage mixtures of exponential families.

Recall that a location is modeled as an area in our approach. To compute the matrix $M^g$ in region $g$, let $P^g = \{P^g_1, P^g_2, \ldots, P^g_n\}$ denote the distributions of scenes estimated as introduced in Section 4.1. The elements $m_{ij}^g$ in matrix $M^g$ are computed as

$$ m_{ij}^g = D_{JS} (P^g_i \parallel P^g_j) $$

(12)

which is the spatial similarity between the corresponding scenes $s_i$ and $s_j$, measured by the Jensen-Shannon divergence.

\(^1\)http://vincentfpgarcia.github.io/jMEF/

5. HIERARCHICAL FRAMEWORK

In the following, we will introduce our proposed automatic video geotagging framework. Technical details of spatial segmentation, visual feature pooling, probability score estimation and integration will be presented.

5.1 Hierarchical Spatial Segmentation

Similar to the work by Kelm et al. [10], we segment the Earth’s surface into regions of different granularities. As shown in Figure 3, Level 1 represents the largest grid with coarse granularity. In this study, we adopt a hierarchy of three levels. The lowest hierarchy level uses a large grid of $36 \times 18$ segments in which the size of the cells is $10 \times 10$, followed by a small grid of $360 \times 180$ segments. Considering that one video can only cover limited aspects of a location, we decide to use small patches of $0.02 \times 0.02$ at the highest hierarchy level instead of directly matching to individual videos.

![Figure 3: The hierarchical spatial segmentation of the Earth’s surface.](image)

5.2 Visual Feature Pooling

Recall that in the feature extraction for videos, we quantized each frame to its most visually nearest scene in the dictionary. This process is referred to as hard coding. Let $X_v = \{x_1^v, x_2^v, \ldots, x_n^v\}$, where $x_i^v$ is the number of frames in video $v$ that have been quantized to scene $s_i$. The final visual representation of the video is achieved by normalizing $X_v$ by its Manhattan norm.

Similarly, the visual features of a region $g$ can be computed by pooling all the frames in it. Let $Y_g = \{y_1^g, y_2^g, \ldots, y_n^g\}$, then $Y_v$ can be computed as $y_i^g = \sum_{v \in g} x_i^v$ where $\{v \in g\}$ represents all the videos located in region $g$. However, to avoid placing too much emphasis on individual videos, we propose to pool the visual features as

$$ y_i^g = \sum_{v \in g} \min\{x_i^v, \text{Scene}_{\text{max}}\} $$

(13)

where $\text{Scene}_{\text{max}}$ is a threshold that limits the influence of individual videos. Note that it may not be a good idea to assign high weights to the scenes that only frequently appear in a limited number of videos, because they are not likely to carry much valuable information of the location. Later we will see in the experiments that the geotagging accuracy can be improved by using a small value of $\text{Scene}_{\text{max}}$. Finally, we normalize $Y_v$ by its Manhattan norm.
5.3 Score Estimation and Integration

So far we have presented the spatial segmentation and visual feature pooling techniques of our proposed framework. Next we will introduce the computation of the probability scores of spatial segments at different hierarchical levels with respect to a given test video. The geotagging process is illustrated in Algorithm 1. Recall that \( P(g|v) \propto P(v|g)P(g) \). The location prior \( P(g) \) models the distribution of where in the world videos are likely to have been taken [6]. In our approach, we compute \( P(v|g) \) and \( P(g) \) with adaptation to region granularities. At Levels 1 and 2 (36 × 18 and 360 × 180 cells) in the hierarchy, \( P(g) \) is estimated as the number of training videos located in a spatial segment divided by the total number of videos in the training set. \( P(v|g) \) is computed using Eq. 3, namely \( P(v|g) = \text{SpatialCon}_{v,g} \cdot \text{VisualSim}_{v,g} \). At Level 3, \( P(g) \) is set to be a small constant and the calculation of \( P(v|g) \) is simplified as \( P(v|g) = \text{VisualSim}_{v,g} \) because the uncertainty of spatial consistency is reduced when using a fine-grained segmentation. Thereafter, \( \text{VisualSim}_{v,g} \) is computed as the cosine similarity between the BoS visual features of video \( v \) and spatial segment \( g \). \( \text{SpatialCon}_{v,g} \) is estimated using Eq. 4, i.e., \( \text{SpatialCon}_{v,g} = X_v \cdot M^g \cdot X_v^T \).

In the existing approaches, the geo-coordinates of training videos are usually considered as the candidates for geotags. Then the best location candidate for a video is selected by iteratively determining the most similar spatial segment at each level. In our proposed framework, two key modifications have been made as follows. First, let \( G_i \) denote the spatial grid of Level \( i \) and \( G \) represent the geotag candidate set. We generalize the approach by combining the probability scores of cells at different hierarchical levels

\[
\text{argmax}_{\hat{g} \in G} \sum_{i=1}^{n} \sum_{g \in G_i} (g \text{ contains } \hat{g}) \cdot \mu^{(n-i)} P(v|g)P(g) \tag{14}
\]

where \( n \) represents the total number of levels, \( i \) denotes the current level, and \( \mu \) is a balancing factor. \( (g \text{ contains } \hat{g}) \) is an indicator function that only selects cells \( g \) containing location \( \hat{g} \). When \( \mu \rightarrow 0 \), this can be interpreted as the process that iteratively determines the most similar spatial segment from the lowest to the highest level. When \( \mu \) becomes larger, the influence of the cells from higher levels becomes bigger on the final geotagging results.

Secondly, as mentioned earlier, we use small patches of 0.02 × 0.02 at the highest hierarchy level instead of matching to individual videos. The location candidates are chosen with adaptation to test videos instead of simply using the centers of patches. For the scenes that appear in a test video, we first find the geo-coordinates of all the relevant frames in the training set. Then, we weigh the geo-coordinates by the saliency score of its corresponding scene with reference to the test video. Thereafter, the weighted geo-coordinates are smoothed by a Gaussian kernel in each patch. Finally, the location candidates for geotagging are formed by the peaks of the Gaussian kernels.

6. EXPERIMENTS

We implemented the hierarchical geotagging framework and evaluated its effectiveness. First, we introduce the dataset and the features we used in the experiments. Next, we analyze the proposed approach based on different model settings. Finally, we compare our proposed model with the existing visual approaches and demonstrate its advantages over its competitors.

6.1 Dataset and Experimental Setup

We evaluated the effectiveness of our approach with the YFCC100M dataset [22], which has also been used in the MediaEval 2014 placing task. The set contains 35,000 geotagged videos in total which have been divided into a training set of 25,000 videos and a test set of 10,000 videos. The training and the test set are mutually exclusive with respect to the users who contributed the media. The geo-coordinates associated with the videos were used as the ground truth. Since the raw geo-coordinates do not always serve as the precise location of a video, we evaluated our method at different accuracy levels of 0.1 km, 1 km, 10 km, 100 km, 1,000 km, and 10,000 km.

Keyframes were extracted at every second of a video. The visual feature we used is GIST [16]. It describes an image based on its spatial envelope, which is a set of perceptual dimensions (naturalness, openness, roughness, ruggedness and expansion) that are related to the shape of space. In the multidimensional space it generates, scenes that share membership in semantic categories (e.g., streets, highways, coasts) are able to be projected closed together.

In our experiments, the vocabulary size is set to 500. Traditionally, k-means clustering is usually applied for dictionary generation. Here we use a simple random selection of scenes due to its high efficiency and similar effectiveness [17]. Thereafter, the keyframes are assigned to the nearest visual scene. Note that here we use hard assignments because it results in a sparse representation. This property greatly accelerates the process of scene distribution modeling. Soft assignments can also be used, but it would take much more time for model building. As we can observe in the experiments, promising results can be achieved by using sparse representation. The evaluation on soft assignment is left as part of the future work.

Algorithm 1 A hierarchical framework for video geotagging

\[
X_v \text{ represents the visual feature of a test video } v \\
Y_g \text{ represents the visual feature of cell } g \\
G_i \text{ is the spatial grid of Level } i \\
G \text{ is the candidate set of geotags} \\
N_g = \text{number of training videos located in cell } g \\
N = \text{total number of training videos} \\
\mu = \text{a balancing factor for score fusion} \\
\text{for } i = 1 \text{ to } 3 \text{ do} \\
\text{if } i = 1 \text{ or } i = 2 \text{ then} \\
\text{else} \\
\text{end if} \\
\text{end for} \\
\text{return } \text{argmax}_{\hat{g} \in G} \sum_{i=1}^{3} \sum_{g \in G_i} (g \text{ contains } \hat{g}) \mu^{(3-i)} P(v|g)P(g)
\]
6.2 Model Analysis

As we mentioned before, our approach segments the globe into grids and models the distribution of scenes in each cell. In our first experiment, we tested two grids that differ in size: a coarse one that segments the Earth’s surface into 36 × 18 cells and a fine one that consists of 360 × 180 cells. Visual features were pooled for each cell as introduced in Section 5.2. A test video was iteratively classified to the spatial segment with the highest similarity score. We varied the threshold Scene_max for feature pooling and reported the classification results in Table 1. We not only showed the classification accuracy, but also computed the geo-location entropy of the correctly classified videos. We first built a histogram, \( D = \{d_1, d_2, \ldots, d_n\} \), regarding the location distribution of the correctly classified videos over the cells. Then we calculated the geo-location entropy, \( Entropy_{geo}(D) \), using the following equation

\[
Entropy_{geo}(D) = -\sum_{i=1}^{n} d_i \log d_i
\]

The video location was interpreted as the same as its most visually similar video in the selected cell. We calculated the average distance error over all the test videos and reported it in the last column.

Table 1: Classification results over different geo-spatial segments.

(a) Spatial grid: 36 × 18 over the globe.

<table>
<thead>
<tr>
<th>Scene_max</th>
<th>Accuracy</th>
<th>Entropy</th>
<th>Avg. Error (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.07%</td>
<td>0.71</td>
<td>5201.60</td>
</tr>
<tr>
<td>5</td>
<td>10.12%</td>
<td>0.73</td>
<td>5194.59</td>
</tr>
<tr>
<td>10</td>
<td>9.99%</td>
<td>0.74</td>
<td>5207.87</td>
</tr>
<tr>
<td>Infinity</td>
<td>9.52%</td>
<td>0.92</td>
<td>5267.19</td>
</tr>
<tr>
<td>Prior</td>
<td>7.59%</td>
<td>0</td>
<td>5047.52</td>
</tr>
</tbody>
</table>

(b) Spatial grid: 360 × 180 over the globe.

<table>
<thead>
<tr>
<th>Scene_max</th>
<th>Accuracy</th>
<th>Entropy</th>
<th>Avg. Error (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.71%</td>
<td>0.61</td>
<td>5934.90</td>
</tr>
<tr>
<td>5</td>
<td>3.66%</td>
<td>0.65</td>
<td>5949.81</td>
</tr>
<tr>
<td>10</td>
<td>3.59%</td>
<td>0.66</td>
<td>5936.82</td>
</tr>
<tr>
<td>Infinity</td>
<td>3.45%</td>
<td>0.74</td>
<td>5887.86</td>
</tr>
<tr>
<td>Prior</td>
<td>3.39%</td>
<td>0</td>
<td>6013.72</td>
</tr>
</tbody>
</table>

As shown, when the value of Scene_max is increased, the classification accuracy drops and the geo-location entropy increases. One possible reason is that we place too much stress on individual videos if we use a large Scene_max. A scene that only frequently appears in one or two videos may not carry much valuable information for geotagging. Therefore, a small value of Scene_max should be adopted in our framework. Note that the last row in the table shows the results computed using only the location priors. All test videos were assigned to the spatial segment with the highest prior, so the geo-location entropy was naturally zero. The classification accuracy and average distance error served as a baseline for comparison.

Next, we fused the probability scores of cells in different hierarchical levels by a parameter \( \mu \) as shown in Eq. 14. In this experiment, \( n = 2 \) and the location candidates are formed by the geo-coordinates of the most visually similar video in every cell of the small grid. Next, we varied the value of \( \mu \) and reported the results in Table 2. The second column shows the classification accuracy over the small grid (360 × 180). The third column reports the average distance error over all the videos in the test set. The last column depicts the ratio of the previous two, which serves as an evaluation metric. A larger ratio indicates a more effective performance, because high precision and low average error are always preferred.

Table 2: Score fusion over different geo-spatial hierarchical levels.

<table>
<thead>
<tr>
<th>( \mu )</th>
<th>Acc. (%)</th>
<th>Err. (10^6 km)</th>
<th>Acc./Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rightarrow 0 )</td>
<td>3.71</td>
<td>5.935</td>
<td>0.625</td>
</tr>
<tr>
<td>0.1</td>
<td>3.68</td>
<td>5.860</td>
<td>0.628</td>
</tr>
<tr>
<td>0.2</td>
<td>3.63</td>
<td>5.761</td>
<td>0.630</td>
</tr>
<tr>
<td>0.3</td>
<td>3.47</td>
<td>5.659</td>
<td>0.613</td>
</tr>
<tr>
<td>0.4</td>
<td>3.36</td>
<td>5.570</td>
<td>0.603</td>
</tr>
<tr>
<td>0.5</td>
<td>3.40</td>
<td>5.534</td>
<td>0.614</td>
</tr>
<tr>
<td>0.6</td>
<td>3.37</td>
<td>5.486</td>
<td>0.614</td>
</tr>
<tr>
<td>0.7</td>
<td>3.29</td>
<td>5.458</td>
<td>0.603</td>
</tr>
<tr>
<td>0.8</td>
<td>3.25</td>
<td>5.425</td>
<td>0.599</td>
</tr>
<tr>
<td>0.9</td>
<td>3.19</td>
<td>5.398</td>
<td>0.591</td>
</tr>
<tr>
<td>1.0</td>
<td>3.25</td>
<td>5.375</td>
<td>0.605</td>
</tr>
</tbody>
</table>

As illustrated, both the classification accuracy and the average distance error decrease when \( \mu \) grows bigger. Generally speaking, by combining the scores we were able to decrease the average distance error. However, it was difficult to maintain the precision at a given small geographical margin of error at the same time. Table 2 gives us some hints on how to choose a good value for \( \mu \) to balance the scores. The ratio of precision and average error remained high when \( \mu \leq 0.2 \), but decreased afterwards. Therefore, we set \( \mu = 0.2 \) in the following experiments. Compared with \( \mu \rightarrow 0 \), the tradeoff is a 2.93% decrease in the average distance error and a 2.16% decrease in the classification accuracy.

To evaluate the effectiveness of scene distribution modeling in video geotagging, we compared our proposed Spatial Relationship Model (SRM) with VRM which computes the score only as the product of the location prior and the cosine similarity between their visual features. Moreover, instead of matching to individual videos in the last step, we used small patches as introduced in Section 5.1. The patch size was set to 0.02 × 0.02. The results are shown in Table 3. As in the previous experiments, we also calculated the classification accuracy and the geo-location entropy of the correctly classified videos over the small grid (360 × 180) and reported the statistics in the last two columns.

We can see from Table 3 that SRM achieved a relatively low average distance error, high accuracy, and high entropy. This indicates the effectiveness of the utilization of scene geo-spatial relationships in geotagging. Moreover, the precision of geotagging at a small margin of error (Radius < 100 km) was greatly improved by using patches instead of individual videos in the last matching step. One possible reason is that the geo-metadata associated with images is usually automatically recorded by GPS, which contains varied errors. Additionally, only one video is unlikely to cover the
Table 3: Evaluation on the utilization of scene distribution modeling in the hierarchical geotagging framework.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision at Radius (km)</th>
<th>Avg. Error (km)</th>
<th>Accuracy</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>1</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>VRM_video</td>
<td>0.01%</td>
<td>0.13%</td>
<td>1.38%</td>
<td>5.21%</td>
</tr>
<tr>
<td>VRM_patch</td>
<td>0.02%</td>
<td>0.48%</td>
<td>2.15%</td>
<td>5.35%</td>
</tr>
<tr>
<td>SRM_video</td>
<td>0.01%</td>
<td>0.12%</td>
<td>1.42%</td>
<td>5.14%</td>
</tr>
<tr>
<td>SRM_patch</td>
<td>0.03%</td>
<td>0.51%</td>
<td>2.24%</td>
<td>5.24%</td>
</tr>
</tbody>
</table>

whole nearby region. Therefore, it would provide a better visual summary of a location when we use small patches rather than videos.

6.3 Comparative Study

To demonstrate the effectiveness of our proposed method, we compared it with the following visual approaches and report the results:

- **Random**: Use the location of a randomly selected video from the development set. It serves as the baseline method and reflects the characteristics of the dataset.

- **ICSI**: A video is represented by the temporal middle keyframe. It searches keyframes in the development set and uses the location of the most visually similar keyframe as the prediction [4].

- **BoS**: Keyframe visual features are pooled into a BoS video representation. The location of the most visually similar video in the development set is used [17].

- **TUB**: The Earth’s surface is hierarchically segmented into cells. Video representations are pooled for each spatial segment using the mean value. A test video is iteratively classified to its nearest neighbor at each level [10].

To provide a fair comparison, we used GIST, as one of the most geographically discriminative features according to Hays and Efros [7], in all methods. The performance can be improved by using more features, but that is not the focus of this study. Kelm et al. [10] detected the national borders using textual information at the first hierarchy level of spatial segments. However, as we purely rely on visual features in this experiment, the national border detection module in TUB was skipped. The similarity score between a cell and a video was calculated as the product of the location prior and the cosine similarity between the visual features. Table 4 reports the geotagging results achieved by different methods. TUBlarge signifies a large grid of 360 x 180 segments of the Earth’s surface, while TUBsmall denotes that the large grid is further halved in each dimension. We also plotted the precision against the geographical margin of error in Figure 4.

The ideas behind ICSI and BoS are the same. Both methods try to estimate the video location as a probability distribution over the globe by matching it to training videos. ICSI chooses to use the visual feature of the middle keyframe as it is likely to be a good representative for the video. BoS has developed a more compact and high-level video representation to improve the geocoding accuracy. However, it still performed relatively poorly for exact location prediction if we directly compare a test video to the ones in the development set. This is likely to be caused by errors in GPS readings and lack of information in individual videos.

![Figure 4: Graph of the precision versus the geographical margin of error for different techniques.](image)

To solve this issue, TUB uses a hierarchical framework and iteratively classifies a test video to its nearest neighbor at each level. As we can see, our proposed approach SRM outperforms all the other methods and achieves the best results overall. SRMcenter interpreted the location as the center of the patch, while SRMpatch modeled the distribution of relevant scenes in the patch and interpreted the location as the center of the Gaussian kernel. These results show that the modeling of the scene distribution really helps to improve the effectiveness of geotagging. By considering the probability of scene co-occurrences in different areas, we were able to improve the accuracy of classification to the correct cell, and thus outperform the existing approaches.

Finally, we compared our method with the visual approach of RECOD [13], proposed for the Placing Task of MediaEval 2014, in Table 5. Most of this year’s participants of the placing task focused more on image geotagging, and did not present any results on the test video dataset. As reported in the working note [13], the RECOD team evaluated their visual approach on the 10,000 test videos with different configurations. RECOD only used one single visual feature HMP, while RECOD2 applied re-ranking to combine multiple visual and audio features including HMP, GIST, and MFCC. As can be observed, our method achieves a better precision at all different accuracy levels, except 0.1 km. Since we currently only use the GIST descriptor, we expect that our method can be further improved by applying a multi-feature fusion approach.

7. CONCLUSIONS

This paper presented a hierarchical visual approach for the automatic prediction of video geotags. First, we discussed the drawbacks of existing methods, and then noted the importance of exploiting the connections between scenes. We modeled the geographic distribution of scenes by GMMs and estimated their spatial similarity by JSD. We implemented a
Table 4: Geotagging results comparison of the proposed and the existing visual approaches.

<table>
<thead>
<tr>
<th>Radius (km)</th>
<th>Random</th>
<th>ICSI</th>
<th>BoS</th>
<th>TUBlarge</th>
<th>TUBsmall</th>
<th>SRMcenter</th>
<th>SRMpatch</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>6</td>
<td>5</td>
<td>15</td>
<td>17</td>
<td>46</td>
<td>51</td>
</tr>
<tr>
<td>10</td>
<td>19</td>
<td>24</td>
<td>39</td>
<td>134</td>
<td>166</td>
<td>223</td>
<td>224</td>
</tr>
<tr>
<td>100</td>
<td>65</td>
<td>108</td>
<td>123</td>
<td>496</td>
<td>506</td>
<td>524</td>
<td>524</td>
</tr>
<tr>
<td>10,000</td>
<td>828</td>
<td>1022</td>
<td>1068</td>
<td>1364</td>
<td>1367</td>
<td>1525</td>
<td>1525</td>
</tr>
<tr>
<td>Avg. Error</td>
<td>6944.52</td>
<td>6680.65</td>
<td>6511.84</td>
<td>5918.26</td>
<td>5916.83</td>
<td>5760.46</td>
<td>5760.33</td>
</tr>
</tbody>
</table>

Table 5: Precision comparison with those reported by participants of the MediaEval 2014 Placing Task.

<table>
<thead>
<tr>
<th>Method</th>
<th>0.1 km</th>
<th>1 km</th>
<th>10 km</th>
<th>100 km</th>
<th>1,000 km</th>
</tr>
</thead>
<tbody>
<tr>
<td>RECOD_2</td>
<td>0.04%</td>
<td>0.18%</td>
<td>1.28%</td>
<td>2.49%</td>
<td>10.58%</td>
</tr>
<tr>
<td>RECOD_3</td>
<td>0.03%</td>
<td>0.14%</td>
<td>1.33%</td>
<td>3.10%</td>
<td>13.93%</td>
</tr>
<tr>
<td>SRM_2</td>
<td>0.03%</td>
<td>0.51%</td>
<td>2.24%</td>
<td>5.24%</td>
<td>15.25%</td>
</tr>
</tbody>
</table>

Acknowledgements

This research has been supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office through the Centre of Social Media Innovations for Communities (COSMIC).

8. REFERENCES