Spatial-Temporal Tag Mining for Automatic Geospatial Video Annotation

YIFANG YIN, National University of Singapore
ZHIJIE SHEN, Hortonworks Inc.
LUMING ZHANG and ROGER ZIMMERMANN, National University of Singapore

Videos are increasingly geotagged and used in practical and powerful GIS applications. However, video search and management operations are typically supported by manual textual annotations, which are subjective and laborious. Therefore, research has been conducted to automate or semi-automate this process. Since a diverse vocabulary for video annotations is of paramount importance towards good search results, this article proposes to leverage crowdsourced data from social multimedia applications that host tags of diverse semantics to build a spatio-temporal tag repository, consequently acting as input to our auto-annotation approach. In particular, to build the tag store, we retrieve the necessary data from several social multimedia applications, mine both the spatial and temporal features of the tags, and then refine and index them accordingly. To better integrate the tag repository, we extend our previous approach by leveraging the temporal characteristics of videos as well. Moreover, we set up additional ranking criteria on the basis of tag similarity, popularity and location bias. Experimental results demonstrate that, by making use of such a tag repository, the generated tags have a wide range of semantics, and the resulting rankings are more consistent with human perception.

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

General Terms: Algorithms, Design, Human Factors

Additional Key Words and Phrases: Video tags, location sensors, mobile videos, geospatial, social media, clustering

ACM Reference Format:
Yifang Yin, Zhijie Shen, Luming Zhang, and Roger Zimmermann. 2014. Spatial-temporal tag mining for automatic geospatial video annotation. ACM Trans. Multimedia Comput. Commun. Appl. 11, 2, Article 29 (December 2014), 21 pages.
DOI: http://dx.doi.org/10.1145/2658981

1. INTRODUCTION

With advances in mobile device technology and network engineering, user-generated videos [Snoek et al. 2011; Rahman et al. 2010] have become very popular in recent years. Many of the videos are created for and used by geography-related applications, such as surveillance or tourism. For example, Webcams.travel is a well-known map-video composite website, presenting worldwide views of tourism, vacations, etc. To search videos from a large corpus, annotation (or tagging) is still one of the most practical and powerful tools [Ames and Naaman 2007]. However, manual annotations are

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).
Authors' addresses: Y. Yin, L. Zhang (corresponding author), and R. Zimmermann, School of Computing, National University of Singapore; email: {yifang, zglumg, rogerz}@comp.nus.edu.sg; Z. Shen, Hortonworks Inc; email: zshen@hortonworks.com.

DOI: http://dx.doi.org/10.1145/2658981

laborious, often ambiguous, and their uneven quality has been well documented [Yan et al. 2008; Suchanek et al. 2008]. In particular, annotating a video is more challenging than annotating an image, because it consists of multiple scenes, where some are easily overlooked. Therefore, researchers have investigated solutions to automate or semi-automate the annotation process. Principally, candidate tags for an image or a video can be inferred from its nearest neighbors based on certain similarity measurements. Some prior solutions only analyzed the visual features of multimedia content, which is very challenging for open domains and usually very compute-intensive [Jain and Sinha 2010]. In recent years, data-driven methods have been suggested which leverage the collective knowledge that resides in some social multimedia applications [Sigurbjörnsson and Van Zwol 2008; Siersdorfer et al. 2009; Wu et al. 2009]. The annotation task can also be addressed by employing relevance models, which are used to estimate the joint distribution of words and images based on a high quality training dataset [Jeon et al. 2003; Monay and Perez 2003; Feng et al. 2004]. With the increasing availability of geotagged images from social sites such as Flickr, geo-aware tag suggestion tools that consider both the geographic context and multimedia content have also been proposed [Abdollahian and Delp 2009; Moxley et al. 2008]. While most of the existing work focuses on entire-video tag suggestions, several techniques have been proposed to localize tags at the shot- or even frame-level granularity [Ballan et al. 2010, 2011; Shen et al. 2011].

In our prior work, we leveraged the geospatial properties of videos and proposed a sensor-rich and data-driven approach to automatically generate tags for them [Arslan Ay et al. 2008; Shen et al. 2011]. This approach does not analyze the visual features, and therefore is particularly effective specifically for geography-oriented videos. This method first models the viewable scenes of the camera as geometric shapes by means of its accompanied sensor data, and then determines the geographic objects that are visible in the video by querying geoinformation databases through the viewable scene descriptions. Subsequently textual information about the visible objects is extracted to serve as tags. However, the data-driven nature implies that the performance of the aforementioned approach significantly depends on the quality of the geoinformation databases used. Previously we built our prototype using geographical information system (GIS) sources, but they can currently still be incomplete. Details are discussed in Section 3.2. In order to enrich the candidate tag repository in our system, this study concentrates on how to screen raw tags from social multimedia websites, build a tag repository, and integrate it with our autoannotation system. The major contributions are as follows.

—We mathematically model the geographic distribution of tags, extract meaningful features from the model, and build both simple and SVM-based classifiers to discover positionable tags. Furthermore, we demonstrate that the simple classifier which does not require manual input can achieve equally good performance compared to the SVM-based approach. Similarly, we model the temporal distribution of positionable tags to mine the duration when they are appropriate to be used.
—To better coalesce with the repository of tags indexed in the spatio-temporal domain, we extend our prior space-only visibility computation algorithm to the spatio-temporally combined domain, mine more information from social multimedia applications to compute tag similarities and popularities, and rescore tags' relevances to videos, achieving a better quality of the generated tags.

The rest of this article is organized as follows. We introduce the related work in Section 2 and review the automatic annotation system proposed in our prior work in Section 3. In Section 4, we explain the method to profile the geographic and the temporal distributions of tags, and index them in the spatio-temporal domain. Next,

in Section 5, we discuss how to acquire more social hints, such as tag similarity and popularity, to rank tags better. Then, we evaluate how well the new data source benefits the autoannotation framework in Section 6. Section 7 concludes.

2. RELATED WORK

Automating or semiautomating the tag annotation process is a popular research topic. A number of studies have proposed state-of-the-art content analysis methods to understand the semantics of multimedia content [Monay and Perez 2003; Feng et al. 2004; Qi et al. 2007]. Alternatively, other studies proposed to leverage crowdsourced Web data, or combine it with visual features [Sigurbjörnsson and Van Zwol 2008; Siersdorfer et al. 2009; Wu et al. 2009]. Social media content, such as videos and images uploaded to YouTube and Flickr, is widely exploited recently. In general, the candidate tags for an image or a video can be suggested by its nearest neighbors. Siersdorfer et al. [2009] proposed to capture the connections between videos using their content redundancy. Ballan et al. [2010] presented a system for video tag suggestions and temporal localization based on the collective knowledge and visual similarity of frames. Several annotation techniques based on relevance models, which are used to estimate the joint distribution of words and images, have also been proposed and have achieved encouraging performance [Jeon et al. 2003; Monay and Perez 2003]. Liu et al. [2007] argued that the performance and scalability of traditional relevance-model-based methods can be limited by the semantic gap and the dependence on training data, and further proposed a dual cross-media relevance model which estimates a joint probability from the expectation over words in a predefined lexicon.

Recently, researchers have investigated the relationship between tags and geocontexts of multimedia content, and used it to suggest tags. Moxley et al. [2008] proposed a tag suggestion method exploiting both content-based analysis and georeferenced information. Given an image to suggest tags, their system queries a number of geographically closeby images, extracts their tags as candidates, and scores them based on their local popularity and the visual similarity between the target image and its neighbors. Abdollahian et al. [2009] proposed a similar method, but it was aimed at video annotation instead. To conduct a visual comparison between the target video and geographically selected images, they segment the video and extract key frames to represent it. These two methods have two limitations compared to ours. First, it is computationally challenging to require a k-nearest neighbor computation for each image/video to suggest tags. Second, without investigating the global distribution of a tag, it cannot reliably be judged whether the tag carries distinguishable semantics in some place even if it frequently appears. For example, *tourism* and *travel* may be popular in places of interest all over the world, to the point where they cannot help users to recall where the image/video was taken.

Larson et al. [2011b] presented three tasks devoted to tagging and geotagging at the MediaEval 2010 benchmarking initiative [Larson et al. 2011a]. MediaEval brings multimedia researchers together to pool research resources and focus efforts on developing solutions for challenging issues facing multimedia indexing and retrieval. Recently, several techniques have been proposed to uncover the relationship between word concepts and geographical regions. Yanai et al. [2009] proposed to use both image region and geolocation entropy to analyze relations between location and visual features. Intagorn and Lerman [2011] proposed that the boundaries of places can be learnt from noisy social annotations. Thomee and Rae [2013] uncovered the colloquial boundaries of locally characterizing regions by innovatively modeling the data using scale-space theory. In the geographical information systems literature, methods for smoothing raw data points to create continuous distributions have been proposed, with the advantage of creating summary statistics that are less sensitive to high-frequency

noise in the data [Brunsdon et al. 2002]. The basic idea is to replace the data points with continuous kernel functions, for instance, Gaussian probability distributions are usually used. Sizov [2010] built a framework named GeoFolk for multi-modal characterization of social media by combining text features with spatial knowledge in order to construct better algorithms for content management, retrieval, and sharing. The method captured the correlations between coordinates and tags by a mixture of latent topics, where a mixture of per-topic Gaussian distributions was adopted.

There exist two studies that are most closely related to ours. Rattenbury et al. [2007] proposed a method for finding the tags that represent places or events. In their method the domain of study is partitioned into segments of some predefined scales, then the tag usage in each segment is analyzed, and the significant segments where the tag is used are identified and judged whether to indicate a place/event or not. Compared to this study, our method does not need to partition the domain, but focuses on street-level positioning and considers the global distribution. Moreover, we analyze the tag similarity to increase the semantic diversity of the generated tags. The other relevant study was proposed by Zhang et al. [2012]. They also investigated the distribution of tags over the temporal and spatial domains, but they used the distributions as features to mine the similarity among tags. Another important difference is that our study demonstrates a novel scenario of using the correlation model of tags and locations, that is, fertilizing the vocabulary for sensor-rich video annotations.

The geocontext of multimedia objects may be used for innovative applications. For example, some studies demonstrated the usage of photos with geocoordinates to create tourism plans [Lu et al. 2010; Gao et al. 2010]. Others used geocoordinates to place the content on a map to facilitate browsing and navigation of images/videos [Toyama et al. 2003; Ahern et al. 2007]. Yin et al. studied the problem of discovering and comparing geographical topics from GPS-associated documents [Yin et al. 2011b] and investigated the problem of mining and ranking trajectory patterns from the uploaded photos with geotags and timestamps [Yin et al. 2011a]. Besides tag annotation and video search, such geographic mining based applications can benefit from the spatio-temporal tag repository we aim to build in this work as well.

3. REVIEW OF THE AUTOMATIC TAG GENERATION SYSTEM

In our prior work, we leveraged the geospatial properties of videos and proposed a sensor-rich and data-driven approach to automatically generate tags for them [Arslan Ay et al. 2008; Shen et al. 2011]. Here we briefly review the key features of this approach and discuss the limitations of the data source we used before.

3.1. System Overview

Since our annotation method is applied to videos enhanced with sensor data, we created special geospatial video recording applications for smartphones. They acquire, process and record location and orientation meta-data along with the video streams. These sensor data are used to model the coverage areas of the video scenes as spatial objects. We introduced a *viewable scene model* (see Figure 2) which describes the scenes visible in the video based on the camera's *field of view* (FOV) [Arslan Ay et al. 2008]. Compared to other video geotagging methods, which usually assign a single geo-coordinate to a whole video [Hong et al. 2011; Tian et al. 2012], ours provides the viewable scenes at frame-level granularity, such that it can enhance the accuracy of video processing based on geocontext.

Next, the annotation process is automated by querying proper data sources using the viewable scene descriptions [Shen et al. 2011]. Figure 1(a) illustrates the framework of our previous autoannotation approach. The method has two major stages. In stage one, the data sources are queried for visible objects in the videos where the objects'

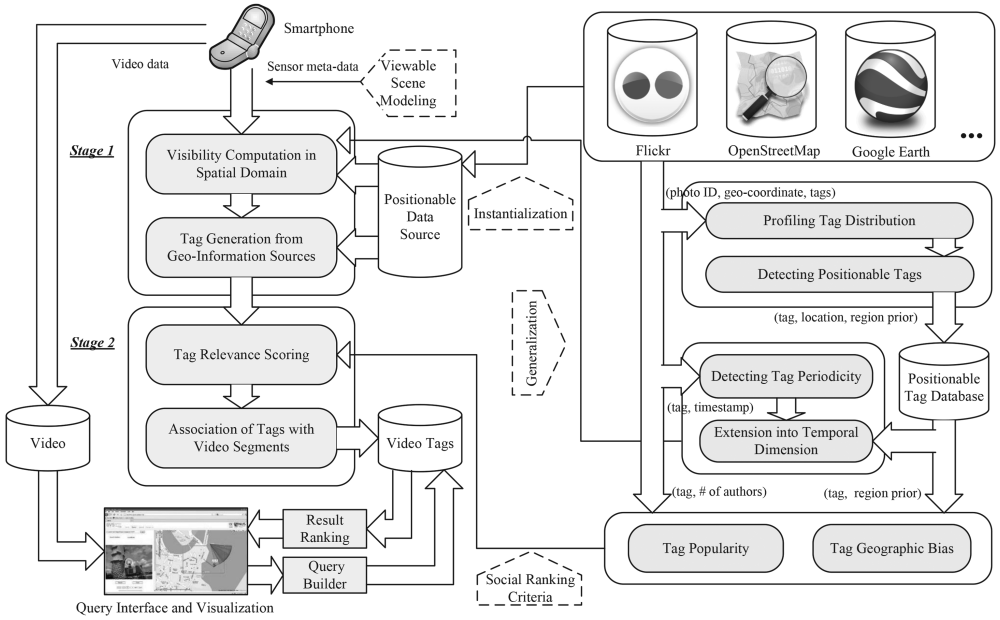


Fig. 1. (a) The architecture of the automatic tag generation framework for sensor-rich outdoor videos, and (b) the process of building a positionable tag repository and interfacing it with the remaining framework.

visibility is calculated through spatial computations. Figure 2 shows the 3D viewable scene model we adopted and Figure 3 illustrates how the visible objects are retrieved for each frame (the visible area is highlighted in blue while the occluded area is highlighted in yellow). In stage two, six relevance criteria are introduced to rank the tags based on their relevance to the videos, which are the *closeness to the FOVScene center*, the *distance to the camera location*, the *horizontally and vertically visible angle ranges*, and the *horizontally and vertically visible percentages*. In our framework, the term *object* is abstract, and can be instantiated in many ways, depending on what the data source is. The only requirement is that an object must be accurately located in some place, such that its relevance to the video can be determined by our viewable scene model. As illustrated in Figure 1(b), this article studies the problem of how to build a rich potisionable tag repository that can be directly applied in the aforementioned annotation system. The basic idea is to mine spatiotemporal tags from social multimedia applications. In the rest of this section, we will first discuss the limitations of the data source we previously used, and then introduce the proposed approaches to incorporate more varied data sources.

3.2. Data Source Limitations

The data-driven nature of the aforementioned approach implies that its performance significantly depends on the quality of the adopted data sources. Previously we built our prototype with the OpenStreetMap[1] (or OSM for short) used as the data source. OSM is a community based map application that can supply detailed information (e.g., names, types, outlines) of numerous geographic objects (or landmarks). However,

[1]www.openstreetmap.org.

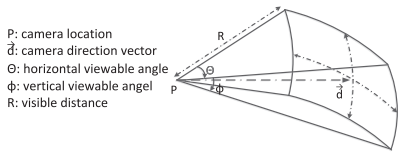


Fig. 2. Illustration of the 3D *FOVScene* model.

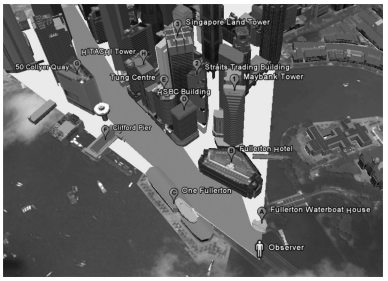


Fig. 3. Illustration of a sample *FOVScene* and the visible objects (or landmarks) which are supplied by Google Earth and determined by conducting geometry computations.

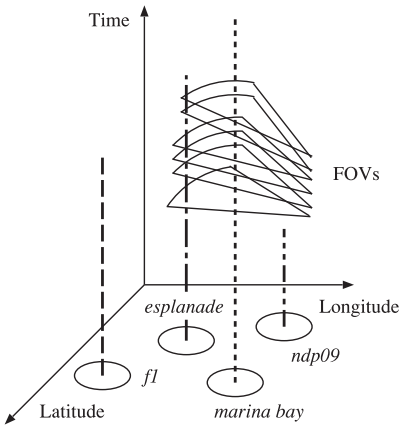


Fig. 4. Conceptual illustration of the placement of tags in the spatio-temporal domain. The dashed lines show the durations of tag usage while the projected circles are the related places of the tags.

its completeness varies in different regions. For instance, the *Merlion* is a popular landmark in the Marina Bay area of Singapore and it was featured in our previous testing videos, but our prototype was unable to recognize it because it is missing in OSM. A more severe problem is that OSM only records landmarks in the physical world, such that the semantics of the generated tags are all within the geospatial domain. In contrast, though we require objects to be associated with some place, they do not necessarily have to be landmarks. Events may also be strongly correlated with a location. For example, the *national day parade*, which is an event, is held in the Marina Bay once a year. Summarily, video tags may miss some important semantics if a system only relies on the data sources of geographic objects. This motivated us to seek more diverse data sources.

3.3. Seeking More Varied Data Sources

We desire that the data sources provide comprehensive information and diverse semantics. However, the objects we investigated in our prior work were only physical entities such as geographic landmarks. In this study we extend the scope and objects can be landmarks, events or other concepts of interest that are positionable in a specific place. One promising source of information is the crowdsourced data available from social multimedia applications, such as Flickr, Picasa and YouTube, where the semantics of images/videos can be acquired by analysing the user-generated tags. Helpfully, the semantics extend beyond the geospatial domain. For example, we retrieved the first $20,000$ images sorted by popularity in the Marina Bay area of Singapore from Flickr and collected their associated tags. Table I lists the top 30 tags and their corresponding semantics, including place, time, event, camera parameters, etc. Meanwhile, these applications support multimedia positioning, that is, images/videos can be assigned a geocoordinate (or geotag). Hence, with images/videos acting as the intermediary, tags and geocoordinates are correlated. This raises the potential that we can discover some tags which are strongly correlated with a specific place. Moreover, the visibility of social tags can be sensitive to time as well (e.g., event tags), which means they are not

Table I. 30 Most Popular Flickr Tags in the Marina Bay Area of Singapore and Their Corresponding Semantics

1–15		16–30	
tag	semantics	tag	semantics
singapore	place†	film	other
f1	event	ndp09	event
marina bay	place	ndpeeps	event
night	time	bw	other
asia	place†	2008	time
canon	camera	skyline	other
esplanade	place	formula 1	event
city	place†	kodak	camera
marina bay sands	place	analogue	other
marina	place‡	travel	other
geotagged	other	analog	other
bay	place‡	black	other
nikon	camera	architecture	other
street	place‡	2009	time
2010	time	river	place‡

applicable to videos that recorded the same place but at different times. This raises the need for us to consider the coverage of a tag in both the spatial and temporal domains.

The data from social multimedia websites is not as organized as that from geoinformation systems, and much of the data are not relevant. To solve this problem, we propose to build a spatiotemporal tag repository that can be directly applied to our autoannotation system, by utilizing the data available from social multimedia applications. As illustrated in Figure 1(b), we collect the tags, the geolocation, and the timestamp associated with multimedia objects. To determine whether a tag is positionable or not, we describe its geographic distribution by a Gaussian mixture model, based on which a classifier is built. Next, we extend the repository into the temporal dimension by predicting the periodicity of each tag. Lastly, we estimate the tag popularity and geographic bias, and integrate these two criteria into the tag relevance ranking. In the next section, we will introduce the methods we adopted to build such a tag repository which is both spatially and temporally indexed (e.g., see Figure 4) by making use of social multimedia applications.

4. POSITIONING SOCIAL TAGS IN THE SPATIO-TEMPORAL DOMAIN

We introduce our approach to make use of social multimedia applications to build a data source of positionable tags, and determine their effective period. First we need to retrieve data from a particular social multimedia information source. In this study, we demonstrate the approach with Flickr. Nevertheless, the method can easily be extended to other similar applications such as Picasa and YouTube, assuming that the applications contain multimedia content associated with tags and geocoordinates. The retrieved data is a collection of multimedia objects, which is formally described as $\mathcal{M} = \{m_i | i = 1, 2, \ldots, k\}$. We let $tags(m)$, $geo(m)$ and $time(m)$, respectively, represent the associated tags, the geocoordinates and the recording time of the object m.

Next, we denote the tag collection of this set of photos with $\mathcal{T} = \bigcup_{\forall m \in \mathcal{M}} tags(m)$, and all the images where a tag $\tau \in \mathcal{T}$ appears as $\mathcal{M}^{(\tau)} = \{m | \tau \in tags(m), \forall m \in \mathcal{M}\}$. Consequently, all the geocoordinates related to a tag can be expressed as $\mathcal{G}^{(\tau)} = \bigcup_{\forall m \in \mathcal{M}^{(\tau)}} geo(m)$, and all the recording times can be similarly formulated as $T^{(\tau)} = \bigcup_{\forall m \in \mathcal{M}^{(\tau)}} time(m)$.

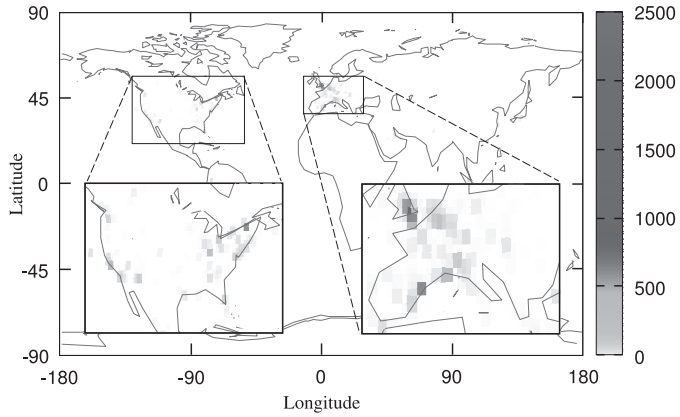


Fig. 5. Illustration of the global distribution of the geocoordinates of tag *f1*.

4.1. Geographically Positioning Social Tags

Importantly, we need to formally define the concept of a *positionable tag*, which is a tag that is strongly correlated to some location at *street level* accuracy. There are two requirements for this. Being strongly correlated indicates that the tag needs to frequently occur in some places but not elsewhere, while reaching street level resolution makes sure that the accuracy level of the location of the tag matches that of our viewable scene model, which is on the order of hundreds of meters. However, not all the tags can meet these two requirements. In Table I, the place tags with a "‡" mark are so general that the distributions of their geocoordinates tend to be relatively uniform. On the other hand, the place tags with a "†" mark are sure to occur more frequently in some places, but the granularity of the places is too coarse to be comparable with our viewable scene model. Note that not just place tags can be positionable. For example, the street course of *f1*, which means the Formula One automobile race, is well defined. Therefore, the first challenge is to determine whether a tag is positionable, and if it is, where the tag is positioned.

To solve this problem, we build a model to describe the distribution of the geocoordinates of a tag, and leverage the expectation maximization algorithm [Dempster et al. 1977] to estimate its parameters. This step is considered as a dimension reduction to some extent. Next, we extract two features from the distribution model and use them to build a classifier to determine whether the tag can be positioned into our area of interest (AOI). Note that since a tag can be positioned anywhere, it is not easy to build a world-wide ground truth to evaluate the performance of our method. Moreover, in many cases, applications may be only interested in some specific places. Hence we properly adapt the original challenge to detect a positionable tag in our predefined AOI. In the remainder of this section, we explain the method in detail.

4.1.1. Profiling Tag Distribution. The geocoordinates of a tag are likely to be unevenly distributed. Figure 5 shows an example of the tag *f1*, where we can observe a number of hot spots (the points in color), indicating the frequent usage of this tag in these regions. To identify where the hot spots are, we construct a high-level mathematical model to describe the distribution of geocoordinates. The basic idea is to replace the geocoordinates with continuous kernel functions to create summary statistics that are less sensitive to high-frequency noise in the data. Intuitively, for a certain tag τ, each hot spot can be modeled with a bivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, where the mean $\mu = \mathrm{E}[\vec{g}] = (\mathrm{E}[lon], \mathrm{E}[lat])^\intercal$ and the covariance matrix $\Sigma = \mathrm{E}[(\vec{g} - \mathrm{E}[\vec{g}])(g - \mathrm{E}[\vec{g}])^\intercal]$

(superscript (τ) is omitted for simplicity). Note that a hot spot is not necessary to be as pronounced, as shown in Figure 5. Assume there are n such normal distributions, and each single geo-coordinate $\vec{g}$ follows either one with the probability γ, where $\sum_{i=1}^{n} \gamma_i = 1$. Hence we can model the distribution of all the geocoordinates as the weighted composite of the n normal distributions, that is,

$$\mathcal{P}_g(\vec{g}|\vec{\gamma}, \vec{\mu}, \vec{\Sigma}) = \sum_{i=1}^{n} \gamma_i \mathcal{N}_i(\vec{g}|\mu_i, \Sigma_i). \tag{1}$$

However, $\vec{\gamma}$, $\vec{\mu}$, and $\vec{\Sigma}$ in Equation (1) are actually unknown variables. We need to estimate them from the set of geocoordinates $\mathcal{G}$ that we obtained. From the probability function, we can derive the likelihood function as

$$\mathcal{L}(\vec{\gamma}, \vec{\mu}, \vec{\Sigma}|\mathcal{G}) = \mathcal{P}_g(\mathcal{G}|\vec{\gamma}, \vec{\mu}, \vec{\Sigma}) = \prod_{i=1}^{\|\mathcal{G}\|} \mathcal{P}_g(\vec{g}_i|\vec{\gamma}, \vec{\mu}, \vec{\Sigma}) \tag{2}$$

or the more convenient log-likelihood function as

$$\hat{l}(\vec{\gamma}, \vec{\mu}, \vec{\Sigma}|\mathcal{G}) = \frac{1}{\|\mathcal{G}\|} \sum_{i=1}^{\|\mathcal{G}\|} \ln \mathcal{P}_g(\vec{g}_i|\vec{\gamma}, \vec{\mu}, \vec{\Sigma}). \tag{3}$$

Consequently, our target can be formalized as

$$\arg\max_{\vec{\gamma}, \vec{\mu}, \vec{\Sigma}} \hat{l}(\vec{\gamma}, \vec{\mu}, \vec{\Sigma}|\mathcal{G}). \tag{4}$$

To find the parameters of our geocoordinate distribution model that maximize the likelihood function, we make use of the well established expectation maximization (EM) algorithm [Dempster et al. 1977]. The EM algorithm is an iterative method: it alternately performs an expectation (E) step, where the expectation of the log-likelihood is evaluated with the current estimations of the parameters, and a maximization (M) step, where parameters is computed to maximize the expected log-likelihood found on the previous E step. One of the issue that has not been clarified is the number of confederate normal distribution n, which needs to be specified during the execution of the EM algorithm. Therefore, we additionally recruit a v-fold cross-validation algorithm[2] to automatically determine how many normal distributions are required to model the distribution of geocoordinates. The general idea is to divide the observed data (or $\mathcal{G}$ here) into v folds. The EM algorithm is respectively applied to the v folds of the training data. The log-likelihood values for all the v folds are averaged into a single metric to measure the stability of our model. At the beginning, the number of normal distributions is set to 1. If the average log-likelihood has been increased, we will correspondingly increment the number of normal distributions by 1 and invoke a new round of cross-validation.

4.1.2. Building a Positionable Tag Classifier. With the distribution characteristics highlighted by the aforementioned model, it is possible to determine whether the tag is positionable in our predefined area of interest (AOI). Intuitively, a tag is considered positioned at the place where a hot spot emerges, and the mean vector $\vec{\mu}$ is consequently regarded as the set of candidate positioning locations. However, not all the hot spots qualify. As is mentioned earlier, the accuracy of the tag position should reach street level, thus the area of the bell-shape of a normal distribution (or the confidence region R_{cr}) should be small enough, such that each mean μ decisively approximates a specific

[2]Electronic Statistics Textbook. StatSoft, Inc. 2011.

location of the tag. The area can be estimated through the covariance matrix Σ, that is, $R_{cr} = var(lon) + 2cov(lon, lat) + var(lat)$. Hence we define the positioning locations of a tag, denoted by $\vec{\mu}'$, as the ones that are subject to $R_{cr} \leq \pi r_0^2$, where r_0 is the threshold of the street-level granularity.

However due to data noise and incompleteness, we found that having one or more positioning locations can not ensure that a tag is positionable. To address this problem, we build a binary classifier $\mathcal{C}$, which takes the information of a tag's positioning locations as input and outputs 1 if it considers the tag to be positionable in the AOI, and 0 otherwise. We employ two features to build the classifier. The first feature $f_1(\tau)$ is the number of positioning locations in the AOI. By definition, $f_1(\tau) = \|\vec{\mu}' \cap \text{AOI}\|$. The second feature $f_2(\tau)$ is the sum of the priors of the positioning locations in the AOI. The prior p is estimated by the Gaussian mixture model. By definition as well, $f_2(\tau) = \sum_{\mu_i \in \vec{\mu}' \cap \text{AOI}} p_i$. We observe that some tags have a hot spot in the AOI but are not widely considered as strongly correlated to the AOI. The reason is that these tags happened to be frequently used by a small number of users in the AOI, such that placing the tags there may not make sense to a majority of users. According to the distribution model, we expect this phenomenon to produce some hot spots with relatively low priors in the AOI. Therefore, we involve a filter to eliminate this hazard. Finally, we can obtain a classifier that is formalized as

$$\mathcal{C}_{\vec{\mu}', \vec{p}}(\tau) = \begin{cases} 1 & \text{if } f_1(\tau) \geq c_0 \wedge f_2(\tau) \geq p_0 \\ 0 & \text{else,} \end{cases} \tag{5}$$

where c_0 and p_0 are predefined thresholds.

One drawback of the above methodology is the need to heuristically assigned thresholds to both features in Equation (5). To overcome this problem, we can leverage a supervised learning algorithm such as SVM [Cristianini and Shawe-Taylor 2000]. First, we select a small set of tags and ask experts to determine whether they are positionable in the AOI. Furthermore, the values of $f_1(\tau)$ and $f_2(\tau)$ of this tag set are computed. Then, we leverage the SVM algorithm to train the classifier $\mathcal{C}_{svm}(\tau)$. One prerequisite of this method is the availability of an annotated training set. For one or a few AOIs, the manual effort is probably manageable, however, for hundreds of AOIs or more, it is too laborious. As a result, if $\mathcal{C}_{\vec{\mu}', \vec{p}}(\tau)$ is not obviously inferior to $\mathcal{C}_{svm}(\tau)$, we prefer the former. This comparison will be further discussed in the evaluation section.

Applying the classification, we now retain a set of tags that are considered as being positionable in the AOI, and denoted as $\mathcal{T}_p = \mathcal{C}(T)$. For each retained tag, we store a tuple ⟨tag, spike center(s), area(s) of the confidence regions, location prior(s)⟩ into a database. It is noteworthy that, (1) a tag may have multiple positioning locations in the AOI according to our classification algorithm, and (2) the database issues are out of the scope of this article, for instance, how to properly index the tuples to accelerate range query processing.

4.1.3. Tag Expansion Based on Geospatial Feature Similarity. As mentioned at the beginning of Section 4.1, there exist tags that are related to places but are difficult to detect because of their uniform distribution or their coarse granularity (e.g., *bay* and *garden*). Fortunately, the meaning of a tag is usually delimited by its geolocation. For example, the tag *garden* is most likely referring to the Gardens by the Bay if we know that it was published near Marina Bay, and thus the location distribution of the tags *garden* and *gardens by the bay* should be highly similar in the AOI of Marina Bay. Based on this observation, we can find the tags that implicitly refer to a specific place by comparing their geospatial distributions in the AOI with the ones of the positionable tags detected by our classifier. Those tags are considered to be geographically positionable as well, and our tag collections are thus further enriched.

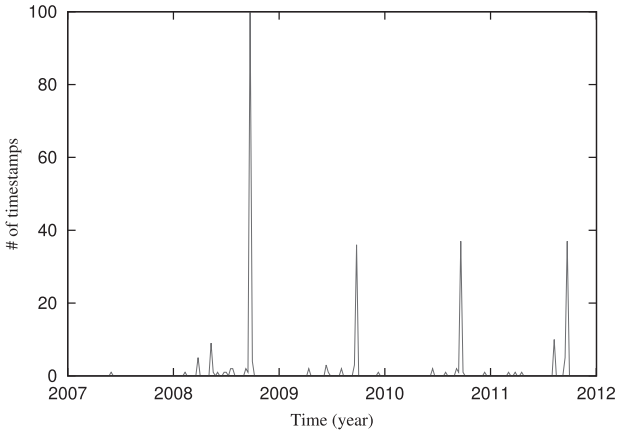


Fig. 6. Illustration of the temporal distribution of the timestamps of tag *f1*.

Zhang et al. [2012] proposed to compute tag geospatial similarity by aggregating tags into geospatial buckets. Here, since we have modeled the distribution of a tag by a mixture of Gaussians, we adopt the *Jensen-Shannon divergence* (JSD) which is a popular method of measuring the similarity between two probability distributions. It is a symmetrized and smoothed version of the *Kullback-Leibler divergence* (KLD), and is defined as

$$D_{JS}(\mathcal{P} \parallel \mathcal{Q}) = \frac{1}{2} D_{KL}(\mathcal{P} \parallel \mathcal{M}) + \frac{1}{2} D_{KL}(\mathcal{Q} \parallel \mathcal{M}), \tag{6}$$

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

$$D_{KL}(\mathcal{P} \parallel \mathcal{Q}) = \int_{-\infty}^{\infty} \ln\left(\frac{p(x)}{q(x)}\right) p(x)\, \mathrm{d}x, \tag{7}$$

where $p(x)$ and $q(x)$ denote the densities of $\mathcal{P}$ and $\mathcal{Q}$. Unfortunately, the KLD between two Gaussian mixture models is not analytically tractable. Here we estimate the KLD between two Gaussian mixture models by the *MonteCarlo* algorithm [Hershey and Olsen 2007]. In our system, we utilized the Java library *jMEF*[3] that can create and manage mixtures of exponential families.

4.2. Temporally Positioning Social Tags

A positionable tag may still not be relevant to some video, even if it is in the coverage area of the video, because its semantics are not valid for the time when the video was captured. It is noteworthy that the semantics of such a tag probably refer to an event. For instance, the tag *ndp09* indicates the National Day Parade held in the area of the Marina Bay on 9 August 2009. While the tag *ndp09* is non-repeatable, the usage of tag *f1* spikes once a year, each time when the Formula One Grand Prix is held in Singapore (e.g., see Figure 6). Therefore, we must estimate the coverage of a tag not only in the spatial but also in the temporal domain.

Currently we only consider the recording times of the photos that are located in the AOI and denote the time set with $T_p^{(\tau)} \subseteq T^{(\tau)}$. Though the data model in the temporal domain is similar to that in the spatial domain, we prefer to use DBSCAN [Ester et al.

[3]http://vincentfpgarcia.github.io/jMEF/.

ALGORITHM 1: Social tags' temporal visible intervals estimation.

Input:
The collection of social tags, $\mathcal{T}$;
The density-based neighbor's reachability parameters, $\epsilon_1 := hour$, $\epsilon_2 := day$ and $\epsilon_3 := month$;
The minimum number of points required to form a cluster, $minPts$;
The threshold parameters, NP, α, $CNum$ and $IC_{\mathcal{I}}$;
Output:
The estimated temporal visible intervals for each tag τ, $\{\mathcal{I}_{vis}^{(\tau)}\}$;
for *each* $\tau \in \mathcal{T}$ **do**
  for $i := 1$ *to* 3 **do**
    $Center, Stddev, noisePerc := DBSCAN(T_p^{(\tau)}, \epsilon_i, MinPts)$;
    if $noisePerc > NP$ *or* $average(Stddev) > \alpha\epsilon_i$ **then** continue;
    for $j := 1$ *to* $\|Center\|$ **do**
      $t_{begin}^j = center_j - stddev_j$
      $t_{end}^j = center_j + stddev_j$
    end
    $\mathcal{I}_i^{(\tau)} = \{[t_{begin}^j, t_{end}^j] | j = 1, 2, \ldots, \|Center\|\}$
    if $\|\mathcal{I}_i^{(\tau)}\| = 1$ **then** /* detect events that happened only once */
      mark tag τ as a single event;
      $\mathcal{I}_{vis}^{(\tau)} := \mathcal{I}_i^{(\tau)}$;
    else if $\|\mathcal{I}_i^{(\tau)}\| \geq CNum$ **then** /* detect periodic events */
      $\mathcal{I}(n), prob := arithProgressionFitting(\mathcal{I}_i^{(\tau)})$
      if $prob \geq IC_{\mathcal{I}}$ **then**
        mark tag τ as a periodic event;
        $\mathcal{I}_{vis}^{(\tau)} := \mathcal{I}(n)$;
      end
    end
    if τ *has been marked as an event* **then** break;
  end
  if τ *is not marked as any event* **then**
    $\mathcal{I}_{vis}^{(\tau)} := any\ time$
  end
end
return $\{\mathcal{I}_{vis}^{(\tau)}\}$;

1996] instead of EM because the density is known beforehand. A repeatable event is expected to occur at a similar hour of different days, or at a similar date/month of different months/years. Therefore, it is very effective to use DBSCAN, which is a density-based clustering algorithm, to discover the time intervals $\mathcal{I}^{(\tau)} = \{i = [t_{begin}, t_{end}]\}$ during which the tag τ is visible in the AOI.

Algorithm 1 sketches the overall procedure to determine a tag's temporal visible intervals. Specifically, we set the level of density reachability ϵ to hour, day and month, respectively, and limit the minimal number of timestamps required to form a cluster to filter small hazard intervals. Next we execute DBSCAN to generate the cluster centers and the standard deviations based on which we further compute the time intervals at different granularity $\mathcal{I}_h^{(\tau)}$, $\mathcal{I}_d^{(\tau)}$, and $\mathcal{I}_m^{(\tau)}$. Subsequently, we analyze the statistics of each tag from the fine-grained to the coarse-grained level to see if a tag's visibility is sensitive to time. We first skip the situations where the timestamps are not well clustered, that is, where the percent of the points that are marked as noise is greater than a threshold NP or where the average standard deviation is α times larger than

the density parameter ϵ. Then, we review the number of clusters generated. If there is only a single time interval (i.e., $\|\mathcal{I}^{(\tau)}\| = 1$), we consider that the tag is representing a single event that is only visible during this time. Otherwise, if the number of clusters generated is greater than a threshold $CNum$, we fit $\mathcal{I}^{(\tau)}$ into an arithmetic progression $\mathcal{I}(n)$. If the fitting achieves a predefined confidence interval $CI_{\mathcal{I}}$, we determine the tag to represent a periodic event that is visible during $\mathcal{I}(0)+k(\mathcal{I}(n)-\mathcal{I}(n-1)), k \in \mathbb{N}$. If a tag is not marked as an event at any granularity, it is considered to be visible at any time.

5. EXTENSION OF THE AUTOANNOTATION APPROACH

Our autoannotation approach can freely incorporate the positionable tag repository. We can compute whether the tags are covered by the viewable scenes of a certain video as used to do it for landmarks. However, we need to extend the visibility computation by adding one more dimension (i.e., time). We compare the timestamp of our *FOVScene* with the temporal visible intervals of the tags. Since determining the visibility of a tag in the time domain is not very computationally complex, we invoke it before performing spatial domain testing, where sophisticated geometry computations are more intense. We make use of the principle location of a tag and assume that its outline is a circle that is congruent with the confidence region. Afterwards, we search for and score any qualified tags for the videos. Finally, an ordered list of tags for each sensor-rich video is obtained. Note that some refinement of the autoannotation approach may lead to a better use of a new data source. Since the tags are obtained from social multimedia applications, crowdsourced data can be leveraged as metrics for tags. These metrics, such as tag popularity and geographic bias, can serve as the criteria to rescore the tags. The popularity of a tag can be estimated by the number of authors who use it. In practice, we select all the multimedia objects in our retrieved data set that are annotated by a specific tag, count the number of unique author IDs, and store them in the database. The priors of the positioning locations of a tag, which is computed when building the Gaussian mixture model, can indicate the tag geographic bias. Next, we present how to use these two measures to rescore the tag relevance.

Our autoannotation system first scores the candidate objects based on their visual relevance to a video. We refer to it as the *baseline score*, $S_b(\tau)$. However, some inherent characteristics of tags are likely to be missing. For instance, the *Esplanade* is a famous landmark in the Marina Bay area of Singapore and one would expect that it attracts more video captures than other, less known structures. However, our experimental system did initially not promote the rank of this tag. Fortunately, social multimedia applications can help to judge the importance of tags. Hence, starting from the baseline score, we propose a promotion score $S_p(\tau)$ to give more credit to important tags.

Recall that the visual relevance of a tag is computed based on the following six criteria: the closeness to the FOVScene center, the distance to the camera location, the horizontally and vertically visible angle ranges, and the horizontally and vertically visible percentages. Since a tag can have multiple positioning locations in the spatio-temporal repository we built, we compute the visual relevance score for each of the positioning locations based on the above six criteria. The baseline score for a tag is subsequently modified to $S_b(\tau) = \sum_i p_i S_b^i(\tau)$, where $S_b^i(\tau)$ represents the visual relevance of the i-th positioning location in the AOI and p_i is the corresponding location prior. Next, we compute the *promotion score* based on the tag popularity which is set to be proportional to the number of authors. Here we prefer widely used tags because they agree with the majority of users' perception and people may be more inclined to use them to search for images/videos as well. Lastly, we linearly combine these two scores for tag relevance ranking:

$$S(\tau) = S_b(\tau) + \omega S_p(\tau), \tag{8}$$

where ω scales the promotion score against the baseline score. As a result, the distinguishable and important tags are promoted, leading to a more appropriate tag ranking mechanism.

6. EVALUATION

We choose Flickr to evaluate the performance of the approach for building a positionable tag repository. The following five AOIs were selected: the Marina Bay in Singapore, the James R. Thompson Center and the Grant Park in Chicago, the Humble Administrator's Garden in China and the Todaiji Temple in Japan. Each of the AOIs was defined as a region of a circle with a radius of 1 km. We compiled the data set from Flickr with the following steps. First, we used the range search API to retrieve the first 20,000 photos taken from 2007 to 2011 in each of the selected AOIs, and ranked according to their popularity. Then, we extracted all the tags used by these seed photos. Thereafter, for each tag, we retrieved at most 20,000 popular photos using it (some tags may not be used by that many photos), and recorded the photo ID, the author ID, the geocoordinates, its accuracy, the recording time and the co-occurrent tags, which make up the dataset. Considering the data noise, we detected and merged duplicate tags by calculating the Levenshtein distance between tags. In the remaining of this section, we demonstrate the accuracy of our positionable tag classification, the accuracy of tag positioning, and the quality of the generated tags by our autoannotation prototype.

6.1. Accuracy of Positionable Tag Classification

To evaluate the performance of our classifier, we selected the 2,500 most frequently used tags (500 per AOI) and invited users to judge whether the tags are associated with a specific place. The tag distributions are modeled as a mixture of Gaussians using Weka [Hall et al. 2009]. Based on this manually annotated ground truth, we first trained and evaluated the performance of the SVM-based classifier $\mathcal{C}_{svm}$. In our implementation, we used LIBSVM [Chang and Lin 2011] to train the classifier, using the number of positioning locations in the AOI (i.e., $f_1(\tau)$), the sum of the priors of the positioning locations in the AOI (i.e., $f_2(\tau)$) and both features, respectively. Only the tags' positioned locations whose confidence region was no larger than the AOI (i.e., $r_0 \leq$ 1km) were considered. The tags with the ground truth were randomly divided into two partitions, that is, a training set and validation set at a rate of 4:1. We ran 40 rounds of classifier training and validation, and in each round, we randomly reselected the training tags to minimize the bias resulting from the training data selection. We use precision and recall as the metrics to evaluate the effectiveness of the classifier. We also report the F1 score, $F_1 = 2 \times \frac{precision \times recall}{precision + recall}$, as it considers both precision and recall.

Table II illustrates the performance of the SVM-based classifier $\mathcal{C}_{svm}$. In general, either the number of positioning locations in the AOI or the sum of the location priors in the AOI is an effective feature, which achieves impressive precision and recall. Using the two features together achieves the best performance in terms of the F1 score. Additionally, we observe that the standard deviations of precision and recall are small, indicating that the performance of the classifiers trained by different data sets is rather stable.

Next, we evaluate the classifier $\mathcal{C}_{\vec{\mu}',\vec{p}}$ based on heuristics. The thresholds are $f_1(\tau) \geq 1$ and $f_2(\tau) \geq 0.6$, with which we obtain a classifier that achieves 0.846 precision and 0.707 recall (see Table III). This indicates that the performance of $\mathcal{C}_{\vec{\mu}',\vec{p}}$ is as good as that of $\mathcal{C}_{svm}$, considering the precision-recall metric. Furthermore, we are interested whether the threshold choice based on our intuition is optimal. Figures 7(a)–(c) describe the performance with respect to the precision-recall metrics over different combinations

Table II. Precision, Recall, and F1 Score Statistics Using the SVM Classifier

	$f_1(\tau)$	$f_2(\tau)$	$f_1(\tau) + f_2(\tau)$
Precision mean	0.735	0.862	0.845
Precision std.	0.032	0.028	0.033
Recall mean	0.826	0.709	0.724
Recall std.	0.028	0.034	0.027
F1 Score mean	0.777	0.778	0.780

Table III. Precision, Recall, and F1 Score Statistics Using the Proposed Classifier by Thresholding

	$accuracy\ level >= 1$	$accuracy\ level >= 14$
Precision	0.846	0.866
Recall	0.707	0.772
F1 Score	0.770	0.816

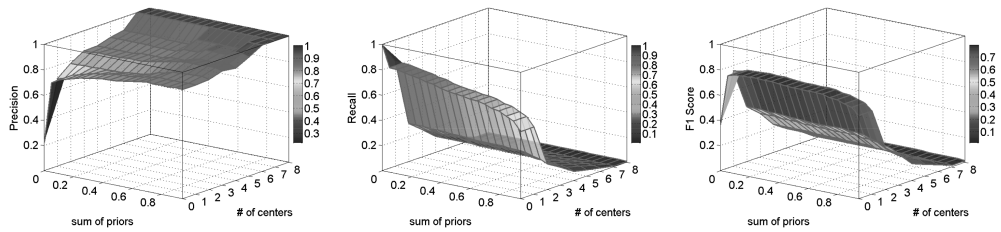


Fig. 7. (a) precision, (b) recall and (c) F1 score under different combinations of the number of centers and the sum of priors thresholds.

of the thresholds of $f_1(\tau)$ and $f_2(\tau)$. Clearly, with an increase of the thresholds, tags are less likely to be considered positionable, such that the precision increases while the recall declines. Considering both the precision and the recall, we observe that the sweet spot is zero or one centers for $f_1(\tau)$ and a not too large percentage for $f_2(\tau)$, where our threshold choices lie. In summary, we can achieve good results with the simple classifier, and need not rely on the SVM-based one that requires manual input.

Additionally, we study the impact of the performance of the geocoordinates on the classification. In practice, the geocoordinates associated with a photo in Flickr may originate from human annotation, or positioning via GPS, cellular base stations or Wi-Fi access points, etc. Different positioning methods have varying accuracy levels. However, we restrictively require each tag to be positionable at some place at street-level accuracy. Therefore, we would expect the accuracy of our classification to be encumbered by inaccurate geocoordinates. Our generic classification approach is blind to the accuracy level of geocoordinates, because the information cannot be assumed to be universally available. Fortunately, Flickr quantifies the accuracy level (from world $\sim$1 to street $\sim$16) and supplies it to API users. Hence, for a subsequent experiment we filtered out the geocoordinates whose accuracy level is below 14 to form the input of our algorithm, and reported the statistics in Table III. By doing so, the classifier achieved 0.866 precision and 0.772 recall with the same threshold settings.

In general, as geotags are collected from crowdsourced media, it is reasonable to assume that the accuracy level of their majority is relatively high. Moreover, the good classification results shown in Figure 7 indicate that our method is capable of filtering out inaccurate data to a certain extent and reflecting the properties of the majority. As

Table IV. F1 Scores Based on Different Settings of NP and α

NP \ α	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
5%	0.55	0.55	0.585	0.585	0.622	0.549	0.536	0.508	0.455
10%	0.651	0.651	0.682	0.682	**0.708**	0.63	0.61	0.581	0.522
15%	0.651	0.651	0.667	0.667	0.694	0.618	0.6	0.571	0.522
20%	0.636	0.636	0.652	0.652	0.68	0.618	0.6	0.571	0.522

Table V. Illustrations of the Estimated Social Tags' Temporal Visibility Intervals

Tags	Center	Std. Deviation	Period	isGeoPositionable
f1	Sep. 28	7 days	Every Year	Yes
2010	Jul. 10, 2010	90 days	—	No
spring	May 1	23 days	Every Year	No
october	Oct. 11	8 days	Every Year	No
christmas	Dec. 29	10 days	Every Year	No
lollapalooza	Aug. 6	2 days	Every Year	Yes
occupy wall street	Oct. 17, 2011	28 days	—	Yes
wall street†	Oct. 15, 2011	7 days	—	Yes
transformers 3†	Aug. 4, 2010	78 days	—	No

pointed out by Hauff [2013], the positional accuracy of the geotag information of Flickr images is highly dependent on the popularity of a landmark. The average distance to the ground truth location is between 11 to 13 meters for images taken at popular landmarks, which is small compared to the size of the viewable scene model we consider.

Next we evaluate the estimation of tags' temporal visibility intervals. We manually annotated tags based on whether they are temporally sensitive or not, and evaluated the effectiveness of Algorithm 1 as a two-class classifier. The dataset was divided into two subsets of equal size, working as the training set and test set, respectively. Now let us recall the input parameters required by the algorithm. $minPts$ denotes the minimum number of points to form a cluster. NP and α are thresholds to skip the situations where the timestamps are not well clustered. $CNum$ and $IC_{\mathcal{I}}$ are parameters for periodic events detection. We set $minPts = 10$, $CNum = 4$, and $IC_{\mathcal{I}} = 0.9$ heuristically, and then tuned NP and α through experiments with the training set. Table IV lists the F1 scores based on different combinations of NP and α values. As shown, the F1 score reaches its maximum when $NP = 10\%$ and $\alpha = 3$, and then decreases on all sides. Therefore, we selected this point as the optimal setting and achieved 0.863 precision and 0.704 recall on the test set. Table V shows some examples of the temporally sensitive tags detected by our algorithm together with the estimated center and standard deviation of their visibility intervals. In general the results are promising. As illustrated, the method is capable of detecting not only the names of single/annual events, but also the tags indicating the time (e.g., month, season, or even year). The last two tags marked by "†" in Table V are examples of false positives generated by our algorithm. Though such tags are usually considered to be visible at all times, on occasion they can be closely related to an event as well. The tag *wall street* is associated with the *Occupy Wall Street* movement which staged a protest event that happened in New York City's Wall Street financial district[4] and the tag *transformers 3* is associated with the filming of the movie "Transformers 3" in Chicago, in 2010. It is difficult to recognize such situations and therefore the algorithm marked them as events as well. We further

[4]The actual event that triggered the hot spot of the tag *wall street* in Chicago is the *Occupy Chicago* collaboration which began on 24 September 2011, in solidarity with the *Occupy Wall Street* protests.

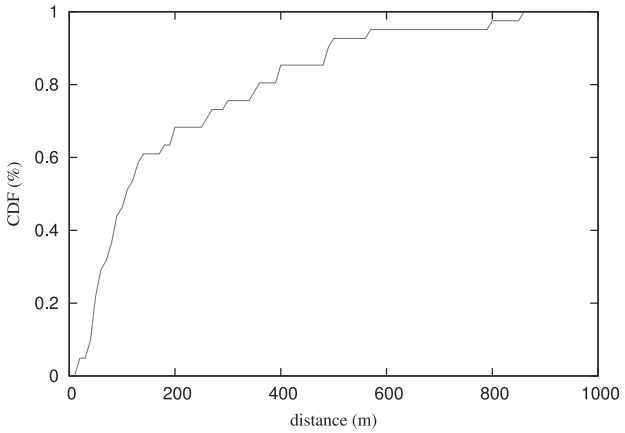


Fig. 8. Cumulative distribution function (CDF) of the distances between the estimated and the real positions.

examined the temporal sensitive tags that were not easily detected and found they mainly included two types: the ones whose deviation was much larger than the density (e.g., *day*, *evening* and *2011*) and the ones that are ambiguous (e.g., *march*).

6.2. Accuracy of Tag Positioning

In the tag classification step, our classifier selected 412 positionable tags that were used for the following evaluation. Recall that we adopt the location of the hot spot covering the highest percentage of geocoordinates as the principle location of a positionable tag. This location is further used by our autoannotation approach to conduct geometry computations and determine the coverage of the tag by a specific video. Therefore, we need to determine whether the estimated locations are accurate enough. Though the classification step ensures that tags are positioned at some locations with street-level accuracy, we need to check whether they are positioned at the locations where they are semantically supposed to belong.

Usually, it is difficult to decide what the correct location of a tag is, except when the tag represents a landmark. We obtained the locations of 41 such positionable tags from Google Map, Wikipedia, etc., to serve as the ground truth, and then computed the distance between the ground truth and our estimated locations. In general, the mean distance is 202 m while the standard deviation is 207 m. In detail, Figure 8 shows the cumulative distribution function of the distances, which are not uniformly distributed. More than 50% of the distances are shorter than 100 m. The absolute values would seem to be still acceptable since the scale of these landmarks is usually at the level of hundreds of meters, and the camera may not be still, but pan across a region.

6.3. Tag Expansion and Ranking

Based on the positionable tags detected by the classifier, we first carried out tag expansion and then supplied these positionable tags to our autoannotation framework, which was equipped with the new features introduced in Section 5. To verify the tag expansion approach, we compute the precisions under different threshold settings and report the statistics in Table VI. The first row lists the results computed based on the true positionable tags that were manually labeled as the ground truth. To eliminate manual work, we carried out the tag expansion based on the positionable tags that were automatically detected, and report the precisions in the second row. As can be seen, both of them achieve the highest precision when the threshold is set to 0.1. Due to error accumulations, the precisions decreased slightly when we utilized the

Table VI.
Precision comparison of tag expansion based on the true positionable tags $geotags_t$ and the automatically detected positionable tags $geotags_d$.

JSD	0.05	0.1	0.15	0.2
Tag expansion based on $geotags_t$	0.815	0.829	0.714	0.632
Tag expansion based on $geotags_d$	0.714	0.778	0.654	0.633

Table VII.
Illustrations of tag expansion. The tag detected is listed together with its nearest positionable neighbor and the *Jensen-Shannon divergence* between them.

Tag	NN	JSD
occupy	nato summit	0.0106
protests	occupy chicago	0.0161
sands	mbs (marina bay sands)	0.0351
skyscraper	downtown chicago	0.0521
bay	marina bay sands	0.0537
downtown	downtown chicago	0.0700
skyway	supertree	0.0949
fountain	grant park	0.0989
bean	attplaza	0.1376
cloud forest	gardens by the bay	0.1507

automatically detected positionable tags. Fortunately, the probability that two random tags are similar in geospatial distribution is low. Compared with the precision of the positionable tag classifier which is 0.846 as reported in Section 6.1, the tag expansion precision 0.778 is compatible and thus can be integrated into our system. Table VII shows some examples of the tags that were expanded.

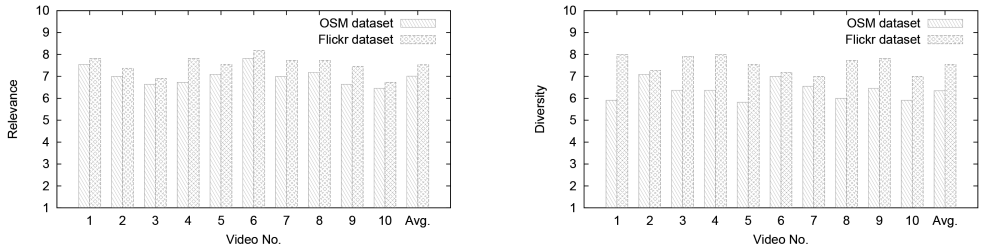
Figure 9 shows two canonical sensor-rich videos we previously captured and the generated tags for each based on different datasets. The recording locations of the video clips were the Marina Bay in Singapore and the Grant Park in Chicago, respectively. For comparison purposes, the first row lists the tags generated using the information extracted from OSM only. The second row of results are generated from the geographically positionable tags that we detected by applying tag classification and expansion. We can observe that the tags in the first row look long, formal and are completely spelled out. In contrast, tags in the second row originate from the Flickr dataset and are more concise and casual. By taking the tags' temporal visibility into consideration, we were able to remove the tags of the National Day Parade and the F1 Grand Prix from the video clip taken near Marina Bay while keeping the tags of the NATO Summit and the Chicago NATO protests for the one taken in Chicago.

To evaluate the effectiveness of our proposed technique, we carried out a user study to capture user preferences regarding the annotation results. We selected ten video clips from different regions around the world. Without loss of generality, we used only the top ten tags generated based on different datasets. 22 volunteers who are familiar with the regions where the videos were taken participated in this user study. They were requested to watch each video carefully and score the tag set based on the following two criteria: (1) the relevance of the generated tags (1 – least, 10 – most), and (2) the diversity of the generated tags (1 – least, 10 – most). Figure 10 shows the results of this user study. As can be seen, the relevance of the tags generated based on either of the datasets is high. The average relevance score achieved by using the Flickr dataset is 7.53, which is higher than the score of 7.01 achieved by using the OSM dataset. The results demonstrate the effectiveness of our proposed techniques to

OSM dataset	Flower Dome, Silver Garden, Marina Bay Sands (Sands SkyPark), Marina Bay Sands, Marina Bay Sands Tower 2, Indian Garden, Dragonfly Lake, The Canopy, Marina Bay Sands Tower 1, Marina Bay Sands Tower 3
Flickr dataset	Marina Bay Sands, Botanical Tower, Marina Bay Reservoir, Marina Bay Gardens, Singapore, Gardens by the Bay, Flower Dome, Marina Bay, Dragonfly, Singapore Flyer

OSM dataset	Grant Park, Chicago Hilton & Towers, Renaissance Blackstone Hotel – Marriott, Columbia College Chicago South Campus Building, Columbia College Chicago 618 S., Michigan Building, Buddy Guy's Legends, Spertus Institute of Jewish Studies, Jones College Prep, Columbia College Chicago Plymouth Court
Flickr dataset	Spertus Institute, Grant Park, Auditorium Theatre, Hilton Chicago, Merle Reskin Theatre, Spertus, NATO Summit, Chicago NATO Protests, Congress Plaza, Chicago Loop.

Fig. 9. Illustration of snapshots of sample videos. The top tags are generated with the proposed auto-tagging system based on different datasets.



(a) (b)

Fig. 10. Comparison of (a) relevance and (b) diversity of the tags generated based on different datasets.

build the spatio-temporal tag repository. In terms of tag diversity, the improvement achieved by using the tag repository we built is even higher. The average diversity scores are 6.35 and 7.55, respectively. As the OpenStreetMap only records landmarks in the physical world, the semantics of the generated tags are all within the geospatial domain. Comparatively, the tags in the spatio-temporal repository we built are not limited to the names of geographic objects but can be any tag that is strongly correlated with a specific place (e.g., the name of an event). Additionally, by applying the tag expansion approach, the semantics of the tags are further enriched. Overall, there is strong evidence that our adaptation algorithms are effective in generating accurate tags with more diverse semantics.

7. CONCLUSIONS

In this article we presented an innovative autoannotation approach for sensor-rich videos, and showed how a positionable tag repository extracted from social multimedia applications can be beneficial. To setup such a repository, we estimated the geographic distribution model of tags, extracted two features from the model, and built two classifiers to detect positionable tags. Furthermore, we profiled their temporal distributions to determine their effective durations. To make better use of the repository, we extended the visibility computation algorithm to the temporal domain, and computed tag similarity, popularity and geographic bias to reorder the tag list. The excellent quality of the generated tags with this overall approach has been confirmed through our evaluation.

In our future work we plan to investigate how to combine tags supplied from heterogeneous data sources, extend our approach to Internet-scale, and popularize our mobile video capturing applications to obtain more sensor-rich videos for evaluation.

REFERENCES

Golnaz Abdollahian and Edward J. Delp. 2009. User generated video annotation using geo-tagged image databases. In *Proceedings of the IEEE International Conference on Multimedia and Expo*.

Shane Ahern, Mor Naaman, Rahul Nair, and Jeannie Hui. 2007. World Explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*.

Morgan Ames and Mor Naaman. 2007. Why We Tag: Motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Sakire Arslan Ay, Roger Zimmermann, and Seon Ho Kim. 2008. Viewable scene modeling for geospatial video search. In *Proceedings of the ACM Multimedia Conference*.

Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Marco Meoni, and Giuseppe Serra. 2010. Tag suggestion and localization in user-generated videos based on social knowledge. In *Proceedings of the 2nd ACM SIGMM Workshop on Social Media*. 3–8.

Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, and Giuseppe Serra. 2011. Enriching and localizing semantic tags in internet videos. In *Proceedings of the ACM Multimedia Conference*. 1541–1544.

C. Brunsdon, A. S. Fotheringham, and M. Charlton. 2002. Geographically weighted summary statistics a framework for localised exploratory data analysis. *Computers Environ. Urban Syst.* 26, 6, 501–524.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*

Nello Cristianini and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines: and Other Kernel-Based Learning Methods*. Cambridge University Press.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Soc. Series B*.

Martin Ester, Hans-peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*.

S. L. Feng, R. Manmatha, and V. Lavrenko. 2004. Multiple Bernoulli relevance models for image and video annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Yue Gao, Jinhui Tang, Richang Hong, Qionghai Dai, Tat S. Chua, and Ramesh Jain. 2010. W2Go: A travel guidance system by automatic landmark ranking. In *Proceedings of the ACM Multimedia Conference*.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explor. Newsl.* 10–18.

Claudia Hauff. 2013. A study on the accuracy of Flickr's geotag data. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 1037–1040.

John R. Hershey and Peder A. Olsen. 2007. Approximating the Kullback Leibler divergence between Gaussian mixture models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 4. 317–320.

Richang Hong, Jinhui Tang, Hung-Khoon Tan, Chong-Wah Ngo, Shuicheng Yan, and Tat-Seng Chua. 2011. Beyond search: Event-driven summarization for web videos. *ACM Trans. Multimedia Comput. Commun. Appl.* 7, 4, 35:1–35:18.

Suradej Intagorn and Kristina Lerman. 2011. Learning boundaries of vague places from noisy annotations. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 425–428.

Ramesh Jain and Pinaki Sinha. 2010. Content without context is meaningless. In *Proceedings of the ACM Multimedia Conference*.

J. Jeon, V. Lavrenko, and R. Manmatha. 2003. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 119–126.

Martha Larson, Maria Eskevich, Roeland Ordelman, Christoph Kofler, Sebastian Schmiedeke, and Gareth J. F Jones. 2011a. Overview of MediaEval 2011 rich speech retrieval task and genre tagging task. In *Proceedings of the Multimedia Benchmark Workshop*.

Martha Larson, Mohammad Soleymani, Pavel Serdyukov, Stevan Rudinac, Christian Wartena, Vanessa Murdock, Gerald Friedland, Roeland Ordelman, and Gareth J. F. Jones. 2011b. Automatic tagging and geotagging in video collections and communities. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. 51:1–51:8.

Jing Liu, Bin Wang, Mingjing Li, Zhiwei Li, Weiying Ma, Hanqing Lu, and Songde Ma. 2007. Dual cross-media relevance model for image annotation. In *Proceedings of the ACM Multimedia Conference*. 605–614.

Xin Lu, Changhu Wang, Jiang M. Yang, Yanwei Pang, and Lei Zhang. 2010. Photo2Trip: Generating travel routes from geo-tagged photos for trip planning. In *Proceedings of the ACM Multimedia Conference*.

Florent Monay and Daniel G. Perez. 2003. On image auto-annotation with latent space models. In *Proceedings of the ACM Multimedia Conference*.

Emily Moxley, Jim Kleban, and B. S. Manjunath. 2008. SpiritTagger: A geo-aware tag suggestion tool mined from Flickr. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*.

Guo J. Qi, Xian S. Hua, Yong Rui, Jinhui Tang, Tao Mei, and Hong J. Zhang. 2007. Correlative multi-label video annotation. In *Proceedings of the ACM Multimedia Conference*.

Abu Saleh Md Mahfujur Rahman, M Anwar Hossain, and Abdulmotaleb El Saddik. 2010. Spatial-geometric approach to physical mobile interaction based on accelerometer and IR sensory data fusion. *ACM Trans. Multimedia Comput. Commun. Appl.* 6, 4, 28:1–28:23.

Tye Rattenbury, Nathaniel Good, and Mor Naaman. 2007. Towards automatic extraction of event and place semantics from Flickr tags. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

Zhijie Shen, Sakire Arslan Ay, Seon Ho Kim, and Roger Zimmermann. 2011. Automatic tag generation and ranking for sensor-rich outdoor videos. In *Proceedings of the ACM Multimedia Conference*.

Stefan Siersdorfer, Jose San Pedro, and Mark Sanderson. 2009. Automatic video tagging using content redundancy. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

Börkur Sigurbjörnsson and Roelof Van Zwol. 2008. Flickr tag recommendation based on collective knowledge. In *Proceedings of the ACM Conference on the World Wide Web*.

Sergej Sizov. 2010. GeoFolk: Latent spatial semantics in Web 2.0 social media. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*. 281–290.

Cees G. M. Snoek, Koen E. A. van de Sande, Xirong Li, Masoud Mazloom, Yu-Gang Jiang, Dennis C. Koelma, and Arnold W. M. Smeulders. 2011. The MediaMill TRECVID 2011 semantic video search engine.

Fabian M. Suchanek, Milan Vojnovic, and Dinan Gunawardena. 2008. Social Tags: Meaning and Suggestions. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*.

Bart Thomee and Adam Rae. 2013. Uncovering locally characterizing regions within geotagged data. In *Proceedings of the ACM Conference on the World Wide Web*. 1285–1296.

Xinmei Tian, Dacheng Tao, and Yong Rui. 2012. Sparse transfer learning for interactive video search reranking. *ACM Trans. Multimedia Comput. Commun. Appl.* 8, 3, 26:1–26:19.

Kentaro Toyama, Ron Logan, and Asta Roseway. 2003. Geographic location tags on digital images. In *Proceedings of the ACM Multimedia Conference*.

Lei Wu, Linjun Yang, Nenghai Yu, and Xian S. Hua. 2009. Learning to tag. In *Proceedings of the ACM Conference on the World Wide Web*.

Rong Yan, Apostol Natsev, and Murray Campbell. 2008. A learning-based hybrid tagging and browsing approach for efficient manual image annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Keiji Yanai, Hidetoshi Kawakubo, and Bingyu Qiu. 2009. A visual analysis of the relationship between word concepts and geographical locations. In *Proceedings of the ACM International Conference on Image and Video Retrieval*.

Zhijun Yin, Liangliang Cao, Jiawei Han, Jiebo Luo, and Thomas S. Huang. 2011a. Diversified trajectory pattern ranking in geo-tagged social media. In *Proceedings of the SIAM International Conference on Data Mining*. 980–991.

Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. 2011b. Geographical topic discovery and comparison. In *Proceedings of the ACM Conference on World Wide Web*. 247–256.

Haipeng Zhang, Mohammed Korayem, Erkang You, and David J. Crandall. 2012. Beyond co-occurrence: Discovering and visualizing tag relationships from geo-spatial and temporal similarities. In *Proceedings of the ACM International Conference on Web Search and Data Mining*.

Received August 2013; revised December 2013, April 2014; accepted June 2014