On Generating Content-Oriented Geo Features for Sensor-Rich Outdoor Video Search

Yifang Yin, Yi Yu, and Roger Zimmermann, *Senior Member, IEEE*

***Abstract*—Advanced technologies in consumer electronics products have enabled individual users to record, share and view videos on mobile devices. With the volume of videos increasing tremendously on the Internet, fast and accurate video search and annotation have become urgent tasks and have attracted much research attention. A good similarity measure is a key component in a video retrieval system. Most of the existing solutions only rely on either the low-level visual features or the surrounding textual annotations. Those approaches often suffer from low recall as they are highly susceptible to changes in viewpoint, illumination, and noisy tags. By leveraging geo-metadata, more reliable and precise search results can be obtained. However, two issues remain challenging: (1) how to quantify the spatial relevance of videos with the visual similarity to generate a pertinent ranking of results according to users' needs, and (2) how to design a compact video representation that supports efficient indexing for fast video retrieval. In this study, we propose a novel video description which consists of (a) determining the geographic coverage of a video based on the camera's field-of-view and a pre-constructed geo-codebook, and (b) fusing video spatial relevance and region-aware visual similarities to achieve a robust video similarity measure. Toward a better encoding of a video's geo-coverage, we construct a geo-codebook by semantically segmenting a map into a collection of coherent regions. To evaluate the proposed technique we developed a video retrieval prototype. Experiments show that our proposed method improves the Mean Average Precision by $4.6\% \sim 10.5\%$, compared with existing approaches.**

***Index Terms*—Video search, feature fusion, geographic coverage, map segmentation, semantic annotation.**

I. Introduction

The ubiquitous availability of smartphones and tablets at affordable prices has encouraged people to engage with the web on the go. Creating, sharing and viewing videos are immensely popular activities with mobile users. Accordingly, user-generated videos have been experiencing unprecedented growth, *e.g.*, more than 100 hours of video are uploaded to YouTube every minute [1]. This high and increasing volume trend has created new challenges for video search. Automatically describing and accurately retrieving video clips from a diverse collection is highly desired.

Traditionally, video search has been conducted by matching query keywords to user generated texts such as titles and tags. However, experience has shown that such metadata are often inaccurate and noisy [2], which leads to an unsatisfactory performance of text-based video search engines. Content-based video reranking [2], [3] is a key technology to address the aforementioned problem, by leveraging content analysis to complement the incompleteness or ambiguity of tags. Unfortunately, this method suffers from the semantic gap [4] that hinders an accurate discovery of video content of interest. To solve this issue, geographic contextual modeling has been investigated recently. Methods have been proposed to judge the relevance of documents based on the textual and spatial similarity with a query [5]. In multimedia, most previous work fuses visual content and geo-context to facilitate image management, whereas little effort has focused on video retrieval. For example, image location information is widely applied for geo-clustering in landmark mining [6], [7], or to create a conjunctive ranking in image annotation and retrieval [8], [9]. Such approaches cannot make full use of the geographic information since in most cases only the camera location is incorporated. For video, in most of the current geo-referenced retrieval systems [10]–[12], clips are ranked purely based on their spatial relevance to the geospatial queries. In this study, we focus on sensor-rich videos where the geographic metadata refers to camera location and orientation. Since such geographic properties are usually automatically recorded using a built-in GPS and compass, we use outdoor videos where the sensor readings are more accurate. We leverage the geographic metadata of videos to improve the performance of text-based and content-based video retrieval techniques. Thus more robust and diverse semantic annotations and similarity search results can be obtained by applying multi-feature fusion.

One issue of the previous fusion approaches is that they utilize the camera location directly. However, such information only captures the camera properties (*e.g.*, photographer location in some street in Paris) rather than the video content (*e.g.*, the Eiffel Tower). This inconsistency motivated us to propose a new content-oriented geo-feature to facilitate video annotation and search. The key components of the approach are illustrated in Fig. 1: the *Hybrid Model Generation* (see Section III) and the *Geo-Codebook Generation* (see Section IV). In the *Hybrid Model Generation* module, a two-level hierarchical model is introduced where multiple cues collaboratively contribute to the video representation. At level one, we generate a geo-histogram which represents the regions that a video covers based on the camera's field-of-view and a pre-defined geo-codebook. Different from the earlier viewable scene model [10] which focuses on individual frames, our proposed model describes the overall geographic coverage of a video. It enables the estimation of spatial relevance between videos through the cosine similarity between the two

Y. Yin, and R. Zimmermann are with the School of Computing, National University of Singapore, 117417, Singapore (e-mail: yifang@comp.nus.edu.sg; rogerz@comp.nus.edu.sg).

Y. Yu is with the Digital Content and Media Sciences Research Division National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, 101-8430, Tokyo, Japan (e-mail: yiyu@nii.ac.jp).

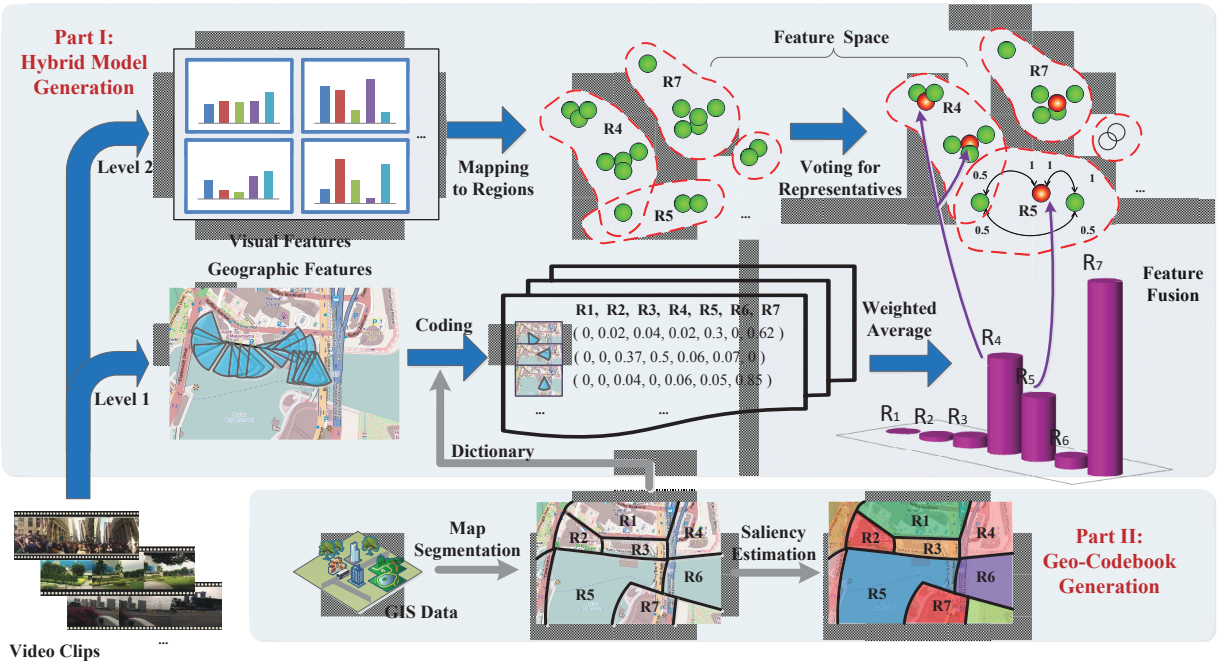


Fig. 1: Illustration of key techniques for geographic and visual feature fusion in our proposed video retrieval system.

corresponding geo-histograms. At level two, we map frames to the regions they capture and select the visually representative ones. Note that in our model, frames are indexed by the regions they capture instead of the camera location. By doing so, geo and visual features are directly connected via regions. Thereafter, we propose a video similarity measure which sums up local similarity scores on a region-by-region basis. Next, toward a better encoding of the geographic coverage in the hybrid model, we present the *Geo-Codebook Generation* module. In this component, we propose an approach that can semantically segment a map into a collection of coherent regions as a geo-codebook. We further quantify the saliency of each region, as humans perceive geographic objects in different areas differently, *e.g.*, a building is more likely to be of interest than a road. Finally, we built a video annotation and retrieval system based on the proposed model. We show that the initial tags generated by querying geo-information services can be enriched by leveraging social multimedia sharing platforms. For evaluation, we carried out a survey to capture user preferences related to the results. Here we summarize the contributions of this study in the following three aspects:

- We propose a novel hybrid model for video representation which generates content-oriented geographic features that can be effectively fused with visual cues to improve the precision of video annotation and search.
- We utilize the information available from the geo-information sources to semantically segment an area into a set of coherent regions, based on which the geographic coverage of a video can be better encoded.
- We have developed a video retrieval prototype based on the proposed video description to demonstrate its effectiveness in retrieving the most relevant search results.

The rest of the paper is organized as follows. We first report the important related work in Section II. The generation of the hybrid model is introduced in Section III, followed by the construction of the geo-codebook introduced in Section IV. Next we present the video annotation and retrieval prototype in Section V. The thorough experimental results in Section VI validate the effectiveness of our system. Section VII concludes and suggests future work.

II. Related Work

Many of the previous text-based video retrieval techniques perform unsatisfactorily due to the mismatch between textual information and video content. To solve this problem, a number of fusion strategies have been developed to improve video retrieval from different modalities [13]. Campbell *et al.* presented a fully automatic retrieval system for speech, visual and semantic modalities [14]. Different types of visual features extracted from keyframes (*e.g.*, color and texture) and text features extracted from speech transcripts were empirically evaluated by experiments for concept detection and video search. To better exploit the underlying relationship between video shots, Liu *et al.* proposed a *PageRank-like* graph-based approach which simultaneously leveraged textual relevancy, semantic concept relevancy, and low-level-feature-based visual similarity in video ranking [3]. Apart from the low-level visual features, more advanced image and video descriptors have been proposed and can be applied in video retrieval systems [15], [16]. Additionally, several multimodal reranking methods have been proposed to improve the initial text search results. Hsu *et al.* proposed a context reranking method by leveraging the contextual information associated with recurrent images or videos over distributed sources [17]. A context graph was constructed where the nodes are videos

and the edges are weighted by multimodal contextual similarities, then the video reranking problem was solved through a random walk on this context graph. Tian *et al.* proposed a content-based reranking technique by formulating video search reranking as a global optimization problem within a Bayesian framework [2]. The conditional prior indicates the ranking score consistency between visually similar samples, and the likelihood reflects the disagreement between the reranked list and the initial one returned by text-based search. However, it is worth emphasizing that none of these methods utilize the geographic metadata which is one of the important kinds of contextual information.

In recent years, geographic metadata has been widely utilized in multimedia mining, annotation and retrieval [18], [19]. Kennedy and Naaman proposed a system that can generate diverse and representative sets of images for landmarks by combining context and content [6]. Crandall *et al.* investigated the problem of organizing a large collection of geotagged photos [7]. Kamahara *et al.* proposed a conjunctive ranking function using both geographic distance and image distance for image retrieval [9]. Liao *et al.* [20] studied geo-aware tag features for image classification. They built tag features by tag propagation from both visual and geo neighbors. For video, Arslan Ay *et al.* proposed to model a camera's field-of-view based on camera position, orientation, viewable angle, and the far visible distance [10]. This viewable scene model was further utilized for efficient video tagging and searching by other work [11], [12], [21], [22]. Arslan Ay *et al.* proposed to rank geo-referenced videos based on three fundamental metrics related to the search area, *i.e.*, the total overlap area, the overlap duration and the accumulation of overlap regions [11]. Zhang *et al.* proposed to calibrate camera location and orientation by registering videos to a mirror 3D world [21], but it requires interactive registration and accurate 3D terrain and building models. Without leveraging the visual features, it is difficult to detect occlusions as this world is not static and we do not have the geo-information of dynamic obstacles such as vehicles.

Unfortunately, few efforts have concentrated on fusing the visual content and the geo-context for a sophisticated video similarity measure. Many of the content-based video retrieval solutions decompose videos into a set of keyframes and define the video similarity based on the pairwise keyframe distances [23], [24]. This prior work usually suffers from low recall rates as the search relies on visual duplication. To better describe video content, modern approaches utilize a set of concepts as intermediate descriptors to facilitate video search [25]–[27]. The concept set is usually general and frequent so as to answer as many queries as possible [28], yet this results in difficulties for the precise interpretation of queries (*e.g.*, queries for a specific building). To overcome these limitations, this study presents a hybrid video representation, based on which precise delimited search results can be obtained. It conjunctively leverages video spatial relevance and local visual similarities in video ranking, and hence it provides excellent support for query-by-example in geospatial video search systems. Experiments show that, based on a geo-referenced video clip or a geotagged image, our proposed system can effectively retrieve the most relevant video clips compared with existing methods.

III. Hybrid Model for Video Representation

While the *viewable scene model* [10] has been adopted for many geo-referenced video applications [11], [21], [22], one fundamental issue is that it describes the camera properties rather than the video content. We argue that content-oriented geo features are highly desired because their consistency with visual clues can make the fusion more seamless. As illustrated in Fig. 2, we propose a novel two-layer model in which frames are indexed by the regions they capture instead of the camera location. Therefore, geo and visual features are directly connected via regions.

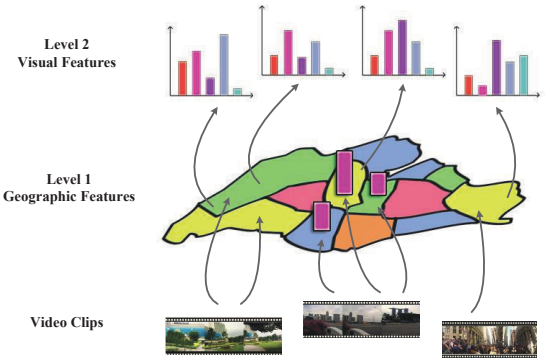


Fig. 2: Illustration of the proposed hybrid model for video representation.

On the first level, this model computes the overall geographic coverage of a video instead of emphasizing individual frames for an efficient spatial relevance measure. On the second level, it indexes frames by regions and selects a number of representative ones based on the visual cues. In the rest of this section, we will first introduce the feature modeling of the proposed two-level video representation and then present a robust video similarity measure based on which more accurate search results can be retrieved.

A. L1: Geographic Coverage Calculation

As introduced earlier, level one aims to capture the overall geographic coverage of a video. To achieve this goal, we pre-segment a map into a set of regions with different saliency values. This is used as a geo-codebook to encode the geo-coverage of a video. The approaches for map segmentation and saliency estimation will be discussed in the next section. The geographic metadata is described by the viewable scene model proposed by Arslan Ay *et al.* [10] (referred to as *FOVScene*). Fig. 3 illustrates the 2-dimensional $FOVScene(P, \vec{d}, \theta, R)$ model overlapping with a geographic region, where ol represents the overlap area, P^c denotes the centroid of overlap ol, and $\vec{d^c}$ is the vector pointing from point P to P^c. These are the important concepts that will be used in the geo-coverage calculation.

To quantify what portion of a region is covered by a frame, we compute the overlap between the camera's *FOVScene* and

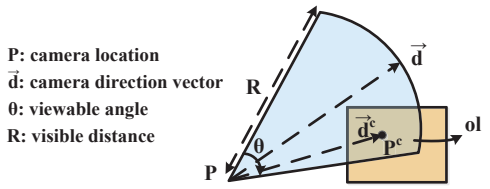


Fig. 3: Illustration of *FOVScene* model in 2D and the concept of geographic overlap.

the regions in the geo-codebook and use the overlap area to emphasize their spatial relevance [11]. As research indicates that people tend to focus on the center of an image [29], we prioritize regions that are close to the camera location and viewing direction [21], [22]. Let ol_{ij} denote the overlap area between region r_i and frame f_j. We assign weights to the regions based on the following three criteria:

- *Normalized area of the overlap*: Considering the regions differ in size, we normalize the area of the overlap $A(ol_{ij})$ by the area of the region $A(r_i)$, that is $\hat{A}(ol_{ij}) = A(ol_{ij})/A(r_i)$.
- *Closeness to the camera location*: We compute the Euclidean distance $D(P^c_{ij}, P_j)$ between the overlap geometry center P^c_{ij} and the camera location P_j, and formulate this criterion as $\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{D(P^c_{ij},P_j)^2}{2\sigma^2})$.
- *Closeness to the viewing direction*: Let $\vec{d^c_{ij}}$ denote the vector pointing from the camera location P_j to the overlap centroid P^c_{ij}. We compute the angular distance $D_\theta(\vec{d^c_{ij}}, \vec{d_j})$ between vector $\vec{d^c_{ij}}$ and the camera direction $\vec{d_j}$, and formulate this criterion as $\frac{1}{\sqrt{2\pi}\sigma_\theta}\exp(-\frac{D_\theta(\vec{d^c_{ij}},\vec{d_j})^2}{2\sigma_\theta^2})$.

Consequently, we compute the weight for region r_i captured in frame f_j using Eq. 1 given below

$$hist^{geo}_i(f_j) = \frac{K_{\sigma,\sigma_\theta}(D(P^c_{ij}, P_j), D_\theta(\vec{d^c_{ij}}, \vec{d_j}))\hat{A}(ol_{ij})}{\sum_k K_{\sigma,\sigma_\theta}(D(P^c_{kj}, P_j), D_\theta(\vec{d^c_{kj}}, \vec{d_j}))\hat{A}(ol_{kj})} \quad (1)$$

where $K_{\sigma,\sigma_\theta}(d, d_\theta) = \frac{1}{2\pi\sigma\sigma_\theta}\exp\left(-\frac{1}{2}(\frac{d^2}{\sigma^2} + \frac{d^2_\theta}{\sigma^2_\theta})\right)$. Since a frame can cover multiple regions, the denominator is a factor that normalizes the sum of the region weights to one.

Subsequently, the geo-coverage histogram for a video v is calculated as the sum of $hist^{geo}_i(f_j)$ using Eq. 2. Since the video segments showing regions with a higher saliency value are more likely to be perceived by humans, we weight the histogram entries by the corresponding region saliency values $saliency(r_i)$, that is:

$$hist^{geo}_i(v) = saliency(r_i)\sum_{f_j \in v} hist^{geo}_i(f_j) \quad (2)$$

Finally, we normalize $hist^{geo}(v)$ by its Euclidean norm:

$$\hat{hist}^{geo}_i(v) = \frac{hist^{geo}_i(v)}{\|hist^{geo}(v)\|_2} \quad (3)$$

Now the geospatial relevance between videos can be efficiently measured as the cosine similarity between the generated geo-histograms, which quantifies the common areas covered by both of the videos:

$$S_g(v_i, v_j) = \sum_k \hat{hist}^{geo}_k(v_i)\hat{hist}^{geo}_k(v_j) \quad (4)$$

Note that for videos where only the GPS location is available in the geo-metadata, it is possible to relax the direction criterion when generating the geo-histograms. We define the geographic area covered by such a frame to be a circle region centered at it with a radius of r. Therefore, Eq. 1 can be reduced to:

$$hist^{geo}_i(f_j) = \frac{K_\sigma(D(P^c_{ij}, P_j))\hat{A}(ol_{ij})}{\sum_k K_\sigma(D(P^c_{kj}, P_j))\hat{A}(ol_{kj})} \quad (5)$$

The regions are weighted based on the first two criteria which are $\hat{A}(ol_i)$ and $K_\sigma(D(P^c_{ij}, P_j)) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{D(P^c_{ij},P_j)^2}{2\sigma^2})$.

The parameters σ and σ_θ in Eq. 1 can be heuristically decided based on Fig. 3. Recall that R denotes the visible distance and θ represents the viewable angle of *FOVScene*. Thus, we set $\sigma = \frac{1}{3}R$ and $\sigma_\theta = \frac{1}{3}\cdot\frac{\theta}{2}$. We use $\frac{\theta}{2}$ because we emphasize the center of *FoVScene*. Subsequently, the visible angle becomes $\frac{\theta}{2}$ to both sides.

B. L2: Representative Visual Features Selection

On the second level, visual features are extracted as the complementary information. In general, it is insufficient to measure video similarity purely based on the common geo-areas covered by both videos because (1) occlusions can occur due to moving objects such as people and vehicles, and (2) the geo-histogram generated on the first level is susceptible to sensor inaccuracy. Therefore, it is highly desired to compare visual features for a more robust similarity measure. The traditional content-based video similarity measures are mostly based on pair-wise keyframe distances. Comparatively, with the prior knowledge of camera location and viewing direction, we can geographically index the frames of a geo-referenced video based on the regions they capture, and compute the local visual similarities in each region.

Since a large number of video frames are near-duplicates, it is necessary to cluster the frames and select their representatives in each region. We build upon an effective lightweight clustering technique called the *reciprocal election approach* proposed by van Leuken *et al.* [30]. The key idea is to let every frame vote for all others. We make adaptations to the voting function to incorporate the frame geo-features. In a video v, let $F = \{f_1, f_2, ..., f_n\}$ denote the set of frames of v that capture the same region r_i. For each frame f_j in F, we rank the others based on their visual similarities to f_j. Particularly, the visual similarity between frames is computed using Eq. 6.

$$W(f_i, f_j) = \exp\left(-\frac{\|f_i - f_j\|^2_2}{\sigma^2_f}\right) \quad (6)$$

Let f denote the k-th nearest neighbor of f_j. The vote f receives from f_j is defined to be $vote(f_j) = hist^{geo}_i(f_j)/k$. A smaller k indicates that the two frames are highly similar and f is a good representative for f_j. A larger $hist^{geo}_i(f_j)$ indicates that f_j is highly relevant to region r_i and it is a salient frame

in set F. Subsequently, the total votes f receives from the others is $\sum_j vote(f_j)$ where f_j is a frame in set F other than f.

After all the frames have cast their votes, the frame with the highest number of votes is selected as the first representative. The cluster around it is formed by those frames whose visual similarity to it exceeds a pre-defined threshold. Next, we exclude the first representative and its cluster members, and select the frame with the highest number of votes in the remaining set as the second representative. This process repeats until the percentage of the remaining frames is less than a threshold.

As the appearance of a region can change among videos, the visual similarity of a region's appearances can be measured based on its representative sets in different videos. To promote visually similar ones in ranking, we present an approach to fuse video spatial relevance with region visual similarity in the following section.

C. Video Similarity Measure

As introduced earlier, the proposed video representation transforms the original per-frame features into per-region features (spatial weight and visual representatives). Subsequently, video similarity can be decomposed as the sum of region feature similarities. As an example, Fig. 4 shows two video clips A and B where a same region, the Marina Bay Sands hotel circled in red, is captured. Recall that the spatial relevance between two videos can be measured as the cosine similarity between the geo-histograms: $\sum_k \hat{hist}_k^{geo}(v_i)\hat{hist}_k^{geo}(v_j)$, that is $0.88 \times 0.6 = 0.528$ between A and B. One issue arises in this case if we measure their similarity purely according to the spatial relevance. That is, the Marina Bay Sands hotel is occluded by trees in B, resulting in a low visual similarity score of 0.43. Furthermore, though the frames circled in blue and yellow show different regions, interestingly they happen to be visually similar.

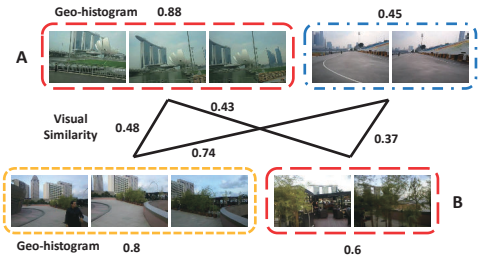


Fig. 4: An example of similarity calculation between two videos.

Without loss of generality, let $w_k^{vis}(v_i, v_j)$ represent the local visual similarity between videos v_i and v_j in terms of region r_k. A small $w_k^{vis}(v_i, v_j)$ indicates that the region's appearances in the two videos are dissimilar, which is possibly caused by unpredictable occlusions, or changes in illumination and viewpoints. Additionally, the geo-metadata error also has an impact on the calculation of the local visual similarity, as it may cause the wrong mappings between frames and regions and therefore affect the selection of the visual representatives. Based on these observations, we penalize such situations by modifying Eq. 4 as follows:

$$Sim(v_i, v_j) = \sum_k w_k^{vis}(v_i, v_j)\hat{hist}_k^{geo}(v_i)\hat{hist}_k^{geo}(v_j) \quad (7)$$

Note that $w_k^{vis}(v_i, v_j)$ can be computed by any existing visual similarity measure [23], [24], [31]. The proposed mechanism conjunctively leverages the geographic coverage similarity and the visual content similarity. $w_k^{vis}(v_i, v_j)$ controls the impact of visual features on the similarity calculation. If $w_k^{vis}(v_i, v_j)$ is set to one under all circumstances, Eq. 7 would become a histogram-based approach which is similar to the one proposed by Arslan Ay *et al.* [11]. The difference is that their approach measures the spatial relevance between a video and a region query, whereas ours focuses on measuring the similarity between two videos. Such methods have the advantages of being highly efficient as the computation is only based on the geographic metadata, but without visual features its performance can degrade due to missed obstacles and occlusions.

On the other hand, the average size of the regions in the geo-codebook controls the impact of geographic features on the similarity calculation. Assume that there is only one region in the geo-codebook which is the entire globe, then Eq. 7 would reduce to one of the existing visual-based similarity measures. In general, better precision can be achieved by using a geo-codebook with finer granularity, as it arranges frames in smaller groups where the visual semantics are more explicit. But considering the errors in GPS and compass readings, a geo-codebook whose granularity is compatible with the size of the *FOVScene* model should be used.

In summary, the proposed model enables efficient spatial relevance calculations between videos as a dot-product of the geo-histograms on the first level and fuses visual clues to promote visually similar ones on the second level. By applying the geographic indexing of frames, our model not only reduces the computational costs, but also excludes noise that exists due to the mismatch between frames from different regions.

IV. Geo-Codebook Generation

The geo-codebook is a key component in the hybrid model generation. Perhaps the simplest way to generate a geo-codebook is to use a grid-based map. However, a grid-based codebook suffers from two drawbacks as shown in Fig. 5. First, geographic objects (*e.g.*, A, B and C) naturally differ in granularity while grid cells are equal-sized. Second, an object (*e.g.*, C) can be separated into multiple cells even if it is smaller than the cell size.

To solve the above two problems, we propose to construct a geo-codebook by a set of coherent regions that cover the map without gaps or overlaps. There are several approaches that can discover the geographic coherent regions by investigating large image collections [32], [33]. However, such techniques cannot be applied for the geo-codebook generation because: (1) the regions discovered are usually not a full coverage of the map, and (2) the granularity of the generated regions is usually too coarse. Alternatively, geo-information services,

Fig. 5: Limitations of a grid-based codebook that cannot satisfactorily capture the diverse granularity of geographic objects.

e.g., OpenStreetMap (OSM), provide information of the geographic objects all over the world. Compared with social image collections, this data source is more detailed and precise based on which a reliable geo-codebook can be generated.

A. Problem Formulation

For a geographic area, we first partition it into a set of square grid cells. Let $G = \{g_i | i = 1, 2, ..., m \times n\}$ denote the set of cells, where m and n represent the number of rows and columns, respectively. Next we retrieve the information of geographic objects in each cell from OSM. Let $O^i = \{o_1^i, o_2^i, ..., o_k^i\}$ represent the object set of grid cell g_i, where k is the total number of objects in it. Each object is represented by a quintuple, $o = \{id, name, tags, footprint, height\}$. A graph $G = (V, E)$ is constructed where the nodes V are grid cells and the edges E are weighted by node similarities. Thereby, the geo-codebook generation can be modeled as a graph clustering problem where each cluster represents a coherent region.

B. Clustering Cells into Coherent Regions

Based on the observation that adjacent similar cells should be merged into the same coherent region, we model the edges in graph G according to the following two criteria, the distance and the similarity between cells, in Eq. 8.

$$e_{ij} = K_\sigma(D(g_i, g_j)) \cdot S(g_i, g_j) \tag{8}$$

where $K_\sigma(d) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{d^2}{2\sigma^2})$; $D(g_i, g_j)$ and $S(g_i, g_j)$ denote the distance and the similarity between grid cells g_i and g_j, respectively.

Intuitively, cells should more likely be merged if they contain one or more common geographic objects. Therefore, we compute $S(g_i, g_j)$ based on the semantic similarity of the geographic objects in them. Recall that the geographic object set in cell g_i is $O^i = \{o_1^i, o_2^i, ..., o_k^i\}$. Further, we assign a weight to each object by measuring the percentage of area it occupies in cell g_i, *i.e.*, $P = \{p_1^i, p_2^i, ..., p_k^i\}$. Thereafter, similarity $S(g_i, g_j)$ is computed as the weighted sum of the pairwise similarity of the geographic objects in grid cells g_i and g_j:

$$S(g_i, g_j) = \sum_{v,w} p_v^i p_w^j S(o_v^i, o_w^j) \tag{9}$$

Recently, Ballatore *et al.* proposed a mechanism to compute the semantic similarity of the OSM geographic classes [34]. They extracted a semantic network from the OSM Wiki website, and computed the tag-to-tag similarity score based on the network topology. As each geographic object can be assigned with multiple tags in OSM, we extend their approach to measure the object-to-object similarity by averaging the corresponding tag-to-tag similarities:

$$S(o_i, o_j) = \begin{cases} 1 & \text{if } o_i.id = o_j.id \\ \bar{S}(t^i, t^j) & \text{else} \end{cases} \tag{10}$$

where t^i and t^j are tags attached with objects o_i and o_j, and $\bar{S}(t^i, t^j)$ denotes the average value of the pairwise tag similarities.

After the graph is constructed, we adopt an effective clustering approach called *Newman and Girvan's Algorithm* [35]. This algorithm avoids the shortcomings of the traditional hierarchical clustering methods by detecting cluster peripheries instead of finding the strongly connected cores. Additionally, it provides a quality measurement called *modularity* which is more effective than empirically chosen thresholds. One issue is that finding a maximum-modularity clustering of a graph is computationally intractable. In our system, we utilized a Java implementation from the project *linloglayout*[1] which used an effective heuristic algorithm for modularity maximization.

Based on the above discussion, semantically coherent regions are obtained, resulting in a descriptive geo-codebook. Therefore, the features encoded in the hybrid model are more explicit and interpretable, leading to a better similarity estimation.

C. Region Saliency Estimation

As aforementioned, the importance of buildings and other geographic objects varies significantly in different areas. For example, landmarks are usually more attractive than ordinary buildings. Therefore, it is necessary to score the regions in the geo-codebook, based on which important objects appearing in a video can get emphasized in the video representation. *Visual saliency* and *social saliency* [36] complement each other in attractiveness estimation. Here we estimate the region saliency according to these two criteria as follows.

Visual Saliency: Higher objects are more likely to draw the attention of the human eye, *e.g.*, a building is more likely to be of interest than a road. Based on this observation, we formulate this criterion as $VS(r) = \sum_i \{p_i \times o_i.height\}$, where o_i represents a geographic object in region r and p_i is the percentage of the area covered by o_i in r.

Social Saliency: This criterion measures the impact of social factors on a region. We collect a set of geotagged images from Flickr, and compute the score for this criterion as $SS(r) = \sum_i K_\sigma(d_i)$, where d_i is the distance between the region center and the location of the i-th image. It can be viewed as the sum of image counts weighted by a Gaussian kernel based on distance.

[1] https://code.google.com/p/linloglayout/

With the popularity of mobile devices and positioning technologies, multimedia contents can be easily recorded together with the location information. Therefore, a number of existing approaches estimate the popularity of a place based on the distribution of geotagged images [7], [37]. Such methods are effective in mining famous landmarks and popular cities. However, they become less descriptive for common regions where not enough photos were taken by people. In order to acquire the ability to effectively rank the less popular regions, we consider the attention-based *visual saliency* additionally. As the traditional attention-based saliency map estimation in computer vision involves intensive computation [38], [39], we adopt a lightweight approach that relies on the attributes of buildings which can be easily acquired from Geographic Information Systems (GIS) [21], [22].

To combine the above two criteria, the saliency of region r is calculated as $saliency(r) = VS(r) + \lambda SS(r)$ where λ is a scaling factor. In famous places around the world, the *Social Saliency* should be the major criterion for scoring. This is because some old buildings, although they may not be tall, are actually very famous and have important societal saliency. Under such circumstances, a larger value of λ should be adopted. On the other hand, in less popular areas, the number of images uploaded to social sharing platforms decreases and the calculation of *Social Saliency* becomes less reliable. As the buildings in such areas are more likely to have relatively equal societal values, the scoring should rely more on the attention-based *Visual Saliency*. Recall that in the geo-coverage calculation, geo-histogram entries are weighted by region saliency scores. Therefore, our proposed hybrid model is able to promote important regions that are more likely to be of interest in the video representation.

V. Video Annotation and Retrieval

We developed a retrieval prototype for sensor-rich videos based on the proposed hybrid model. Besides traditional geospatial queries, our system supports query-by-example where the search results are ranked using the similarity measure introduced in Section III-C. Videos in our system are indexed using inverted files according to the geographic regions they cover. Therefore, only the geo-relevant videos will be processed for similarity computations for efficiency.

Video text annotation is a useful and powerful feature to facilitate video search and browsing in many social media and web applications. However, the majority of tags assigned to videos still come from the manual annotations, which are not only highly time-consuming but also often inaccurate and incomplete. To handle sensor-rich videos, Shen *et al.* [22] proposed to detect the visible geographic objects, the textual cues of which are extracted to serve as geotags. However, this method has one limitation that only geographically related textual tags can be retrieved. To enrich the semantics of the tags generated by this approach, state-of-the-art tag suggestion and refinement techniques can be applied with the help of social multimedia sharing services such as Flickr. More specifically, the names of the visible geographic objects computed based on the geo-metadata will serve as the initial tags. Thereafter, tag refinement techniques based on social knowledge will be performed to enrich the initial tags by utilizing a Flickr image set.

Let $F = \{f_1, f_2, ..., f_n\}$ denote the representative keyframes of a video and $T = \{t_1, t_2, ..., t_m\}$ denote the list of tag candidates for annotation. In the following we briefly describe two data driven approaches that can be used to improve the initial tags generated by Shen's method. The refined scores of tags are represented by an $n \times m$ matrix L where l_{ij} is the confidence score of frame f_i associated with tag t_j.

A. Neighborhood similarity measure: Wang et al. [40]

Recently, graph-based semi-supervised learning has gained much attention in this domain. In the regularization frameworks such as *LLGC* [41], one of the crucial factors is the estimation of the pairwise similarity between images. Traditionally, the similarity between two samples is estimated based on the Euclidean distance between them. However according to Wang *et al.* [40], this distance-based similarity measure may lead to high classification error rates. Therefore, they proposed a novel neighborhood similarity measure which outperformed the Euclidean distance in video annotation as they pointed out that the similarity between samples is not merely related to their distance but also related to the distribution of surrounding samples and labels.

B. Tag suggestion and localization: Ballan et al. [42]

As we mentioned earlier, the initial tags associated with frames in our framework are the names of the visible geographic objects detected from OpenStreetMap. In Ballan's approach [42], they used the initial tags as queries to retrieve Flickr images, based on which visual neighborhoods of the keyframes were created. The union of the tags associated with the images in the neighborhood was considered as the candidates to annotate a keyframe. These tag candidates were next ranked based on the relevance score computed as the count of a tag t in the visual neighborhood of the keyframe minus the prior frequency of t [43], [44].

By applying the above techniques, we were able to get the refined score matrix L. Based on L, the relevance of tags to a video, L_v, is computed as follows,

$$L_v = \sum_{i,j} \alpha_{ij} \hat{hist}_i^{geo}(v) L_j \tag{11}$$

where L_j is the j-th row of matrix L. Let $\mathcal{R}_i$ denote the set of representative frames of region r_i in video v. Then, α_{ij} is defined as:

$$\alpha_{ij} = \begin{cases} \frac{1}{|\mathcal{R}_i|} & \text{if } f_j \in \mathcal{R}_i \\ 0 & \text{else} \end{cases} \tag{12}$$

We will see later in the experiments that the utilization of Flickr images can not only promote the popular geotags (*e.g.*, names of landmarks), but also greatly diversify the semantics of the candidate tags.

VI. Evaluation

We implemented a video search prototype and evaluated its effectiveness. We proceed in three steps. The first part shows two examples of the geo-codebook generation. The second part evaluates the performance of the proposed model in video retrieval. The third part reports the results of a user study that demonstrates the advantages of the proposed semantic annotation approach over its competitors.

A. Experimental Setup

We evaluated our proposed approach on the publicly available geo-referenced video dataset from the GeoVid[2] website. It hosts more than 1,500 videos recorded by smartphones from all over the world. The videos and their corresponding geographic metadata can be retrieved via the provided web APIs[3].

Besides the above dataset, another dataset comprising 15,616 geotagged images was collected from Flickr by performing keyword-based search. Two types of tags were used as the query keywords: (1) the textual information of the geographic objects and (2) 25 popular concepts including airport, animal, birds, boat, bridge, buildings, cityscape, clouds, college, crowd, dancing, flowers, food, garden, grass, lake, person, plants, sky, street, sunset, temple, tree, vehicle, and water. This image dataset was used in both the region saliency estimation and the video semantic annotation.

For each of the frames and images, we extracted the following three low-level visual features in our experiments:

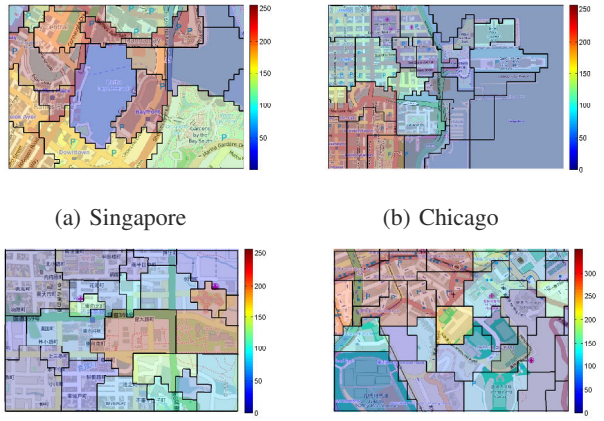
- *48-D Gabor Wavelet Texture:* Texture features extracted at four scales and six orientations using a Gabor wavelet decomposition [45].
- *225-D Block-Wise Color Moments:* The first (mean), the second (variance) and the third order (skewness) color moments in HSV space extracted over 5×5 fixed grid partitions [46].
- *512-D Gist Descriptor:* The spatial structure of an image described by global features derived from the spatial envelope [47].

These features are used for visual similarity measurement.

B. Geo-Codebook Generation

In our implementation, the geographic information of objects was collected from the OpenStreetMap (OSM) which is a geo-information service that provides editable maps of the world. We recorded the name, the tags, and the footprint of each object. However, for buildings described in the OSM, interestingly the height attribute is mostly not available. To solve this problem, we collected the building heights from EMPORIS[4], a real estate data mining company collecting and publishing data and photographs of buildings worldwide. In Singapore for example, it has records of 6,915 buildings, 321 of which have the height information. For the rest where the height of the building is not available, we estimate based on other clues, *e.g.*, the number of storeys.

[2]http://geovid.org/
[3]http://api.geovid.org
[4]http://www.emporis.com/



(a) Singapore (b) Chicago

(c) Japan (d) Hong Kong

Fig. 6: Examples of the generated geo-codebook in different areas around the world.

Fig. 6 presents examples of the generated geo-codebook in four different areas, namely Singapore, Chicago, Japan, and Hong Kong. The cell length of the grids was set to 50 m and the parameter σ in Eq. 8 was set to 65 m ($\sigma = 1.3 \times 50 = 65$ m). This parameter σ controls the connections between adjacent cells. If a large value is used, a cell will be bonded with its neighbors more tightly and therefore result in a coarse geo-codebook. Conversely, a small value of σ will result in a fine-grained geo-codebook. Additionally in Fig. 6, the colors indicate the estimated saliency for each region and the scaling parameter λ was empirically set to 0.4. Compared with the grid-based codebook in Fig. 5, we can see that this model successfully captures the diverse granularity of different geographic objects, and the estimated saliency is also consistent with human perception. Let us take Fig. 6(a) as an example since it shows the same area as in Fig. 5. In the center of the picture, we can see that the shape of Marina Bay (Object A in Fig. 5) is well captured by the geo-codebook. The building on its right (Object B in Fig. 5) is the most famous Marina Bay Sands hotel. Other salient regions marked in red are mainly the popular landmarks including the Singapore Flyer, the Esplanade, the Singapore River, and the financial district. In Fig. 6(b), the salient regions belong to the Loop which is the central business district of Chicago. In Figs. 6(c) and 6(d), the salient regions are the Kofukuji Temple and the Time Square (Hong Kong), respectively.

Note that the current geo-codebook was generated within a city. For large-scale video datasets, our method can be easily scaled up by using a hierarchy: (1) segment the Earth surface into countries and cities, (2) generate geo-codebooks within cities, and (3) index videos using the generated geo-codebooks in various cities.

C. Evaluation on Video Retrieval

To evaluate the effectiveness of our proposed model in video retrieval, we collected a total of 423 videos, ranging from 21 to 523 s in duration. Considering the geo-referenced videos in GeoVid are unevenly distributed, we selected popular

regions (*e.g.*, Singapore and Chicago) where the videos are more concentrated to collect the dataset for experiments. The videos were further segmented into 1,656 shots, each of which are about 30 s in duration. Furthermore, we selected 50 video clips and 30 Flickr images (see Fig. 7) as queries. The selection criterion is that they contain some recognizable places and landmarks which are more likely to be of interest.

Fig. 7: Illustrations of geotagged Flickr images used as queries.

In our implementation, we adopt the method proposed by Cheung *et al.* [23] to measure the distance based on visual clues. Thereby, $w_k^{vis}(v_i, v_j)$ in Eq. 7 is computed as:

$$w_k^{vis}(v_i, v_j) = \exp\left(-\frac{D_k^{vis}(v_i, v_j)}{\sigma}\right) \tag{13}$$

$$\begin{aligned} D_k^{vis}(v_i, v_j) = &\frac{\sum_{f_v \in \mathcal{R}_k(v_i)} \left(\min_{f_w \in \mathcal{R}_k(v_j)} \|f_v - f_w\|_2\right)}{|\mathcal{R}_k(v_i)| + |\mathcal{R}_k(v_j)|} \\ &+ \frac{\sum_{f_w \in \mathcal{R}_k(v_j)} \left(\min_{f_v \in \mathcal{R}_k(v_i)} \|f_v - f_w\|_2\right)}{|\mathcal{R}_k(v_i)| + |\mathcal{R}_k(v_j)|} \end{aligned} \tag{14}$$

where $\mathcal{R}_k(v)$ denotes the set of representative frames of region r_k in video v, $|\mathcal{R}_k(v)|$ represents its size, and $D_k^{vis}(v_i, v_j)$ is the visual distance between the two sets of frames, $\mathcal{R}_k(v_i)$ and $\mathcal{R}_k(v_j)$. As can be seen, we first compute the local visual distance $D_k^{vis}(v_i, v_j)$ as the average distance between the closest matched frames using Cheung *et al.*'s method [23]. Then, we use a Gaussian kernel to acquire the local visual similarity score, which is $w_k^{vis}(v_i, v_j)$.

For the hybrid model generation, we empirically set $\sigma = \frac{R}{3}$ and $\sigma_\theta = \frac{\theta}{6}$ in Eq. 1, where R and θ denote the visible distance and the viewable angle of the *FOVScene* model illustrated in Fig. 3.

1) Effectiveness comparison: To evaluate the effectiveness of our proposed region-aware video similarity measure, here we compared the following four methods and reported the results:

- *GEO*: It ranks videos based on the geospatial relevance using Eq. 4.
- *CRLF*: It filters the collection based on location, and then ranks the remaining based on visual similarity [48].
- *CRGV*: It ranks the collection based on a conjunctive function using both geographic distance and visual distance [9].
- *RASM*: It ranks videos based on the proposed region-aware similarity measure using Eq. 7.
- *OB*: A visual approach based on the state-of-the-art ObjectBank image descriptor. It represents an image based on its response to a large number of pre-trained object detectors [16].
- *BoS*: A visual approach based on the state-of-the-art Bag-of-Scene video representation. It generates a compact descriptor based on a dictionary of scenes, each of which represents a semantic concept [15].

As the existing work [9], [48] built their model using only GPS, to make it a fair comparison we generated the geo-histograms using Eq. 5 in this experiment. Later we will discuss how the performance can be further improved when camera direction is also available in the geo-metadata. For each of the queries, we examined the results and plotted the average precision at n (P@n) in Fig. 8. We also compared the methods based on the Mean Average Precision (MAP) measure which is reported in Table I.

TABLE I: MAP comparison of the proposed and the existing fusion methods.

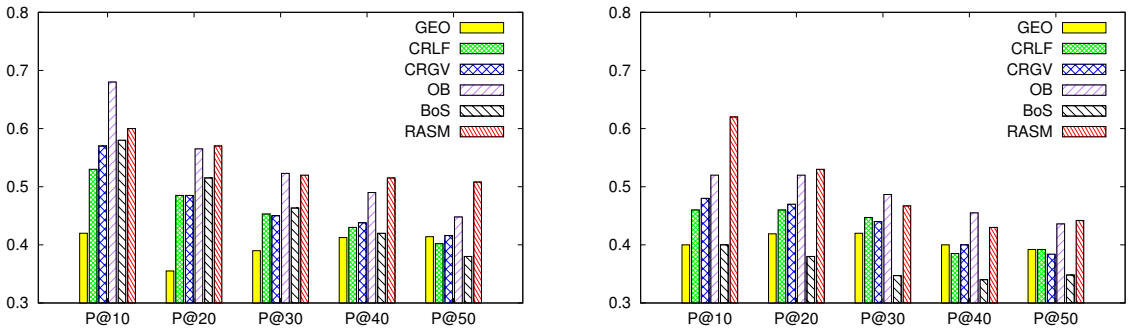
Method	*GEO*	*CRLF*	*CRGV*	*OB*	*BoS*	*RASM*
By-video	39.6%	40.6%	41.2%	44.6%	38.7%	49.2%
By-image	38.9%	39.7%	39.8%	41.7%	34.6%	44.2%

GEO serves as a baseline method because it ranks videos based only on the geo-metadata. *CRLF* and *CRGV* outperformed the baseline method by integrating the visual clues. One issue is that these fusion approaches utilized the camera location directly. However, such information only captures the camera properties rather than the video content. This inconsistency between geo and visual features limited the effectiveness of such approaches. Additionally, we carried out experiments using the state-of-the-art visual features *OB* and *BoS* for comparison. *OB* is an object-level image descriptor which is generated based on pre-trained object detectors. It increased the MAP compared with methods *CRLF* and *CRGV* where the low-level visual features were adopted. However, due to the high dimensionality of the ObjectBank descriptor, it has the drawback of being time-consuming in feature extraction and similarity calculation. The time complexity of each method will be compared in Section VI-C4. In contrast, *BoS* is a high-level compact video descriptor. In this experiment, we used a dictionary of 500 concept scenes and soft coding technique. The *BoS* descriptor represents a video segment using a single vector. Therefore, it is highly efficient in computing the similarity score between videos (see Table VI). However, it might be difficult to maintain a high MAP at the same time. As can be seen, our hybrid model *RASM* achieved the best results overall. It improved the MAP by 4.6% and 10.5% compared to *OB* and *BoS*. Our model generates the geo-coverage of a video which is a content-oriented geo-feature. Good performances can be achieved by fusing only with the low-level visual features. Moreover, our proposed model also works well with more advanced visual features such as *OB* and *BoS*. As reported in Table II, our fusion technique can improve the MAP by as much as 7.7% compared with the original content-based approaches.

TABLE II: MAP comparison of fusion with OB and BoS.

Method	*OB*	*RASM$_{OB}$*	*BoS*	*RASM$_{BoS}$*
By-video	44.6%	51.9%	38.7%	46.4%
By-image	41.7%	48.4%	34.6%	42.1%

As a final point, our model can make use of multiple geo-features in the metadata, while how the camera direction can



(a) Query-by-video. (b) Query-by-image.

Fig. 8: P@n comparison of the proposed and the existing fusion methods.

be utilized in other methods remains unknown.

2) Geo-metadata availability: Next, we studied how the retrieval performance varied when the geo-metadata was available at different levels. The comparison of average P@n is illustrated in Fig. 9, and the MAP statistics are reported in Table III. The subscript indicates which geo-metadata was used in the geo-coverage modeling.

TABLE III: MAP comparison based on different availability of geo-metadata.

Method	GEO_{gps}	$RASM_{gps}$	GEO_{fov}	$RASM_{fov}$
Query-by-video	39.6%	49.2%	66.9%	71.8%
Query-by-image	38.9%	44.2%	48.4%	53.2%

As can be seen, the effectiveness of both *GEO* and *RASM* was greatly improved by utilizing camera direction. It indicates the importance of camera orientation in video content analysis, but unfortunately compass record is still only available in the minority of multimedia documents. Such geo-restrictions can greatly help reduce the semantic gap between the low-level visual features and the high-level semantic concepts. For query-by-video, $RASM_{fov}$ improved the MAP by 22.6% compared to $RASM_{gps}$. For query-by-image, the increments were 9.0%. *RASM* is more robust than *GEO* because its similarity measure is more tolerant to dynamic obstacles and geo-metadata errors by analyzing the visual clues.

In terms of geo-metadata, social sharing platforms such as Flickr provide an accuracy level of geotags associated with photos. Therefore users can avoid using images with inaccurate geotags as queries. As pointed out by Hauff [49], the positional accuracy of the geotag information of Flickr images is highly dependent on the popularity of the venue. The average distance to the ground truth location is between $11 - 13$ meters for images taken at popular venues, which is small compared to the size of the viewable scene model that we consider. Moreover, the good retrieval results shown in Figs. 8 and 9 indicate that our method is robust within a certain range of geotag errors.

3) Step-By-Step Model Justification: The proposed video similarity measure includes two main components: geospatial relevance calculation and multi-feature fusion. To demonstrate the effectiveness of our proposed approach in each step, we replace our method by a functionally reduced counterpart and compare the corresponding retrieval performance.

- The geo-codebook generation is a key component in the first step. We use it to encode the geo-histograms, based on which the geospatial relevance between videos is computed. To illustrate its effectiveness, we replace it by a grid-based approach. Each region in the grid-based codebook is a square area that has a side length of 300 m.
- To justify the effectiveness of the region-aware fusion approach illustrated in Eq. 7, we compare it with the late fusion method [5]. The similarity is estimated as $S = \frac{1}{2}(S_g + S_v)$, where S_g and S_v denote the geospatial relevance and visual similarity, respectively. As shown in Fig. 4, additional noise can be introduced by late fusion due to the mismatch between visual features from different regions.

TABLE IV: Mean average precision decrement.

Query Type	Query-by-video	Query-by-image
geo-codebook→grid map	-2.7%	-2.1%
region-aware→late fusion	-4.3%	-4.4%

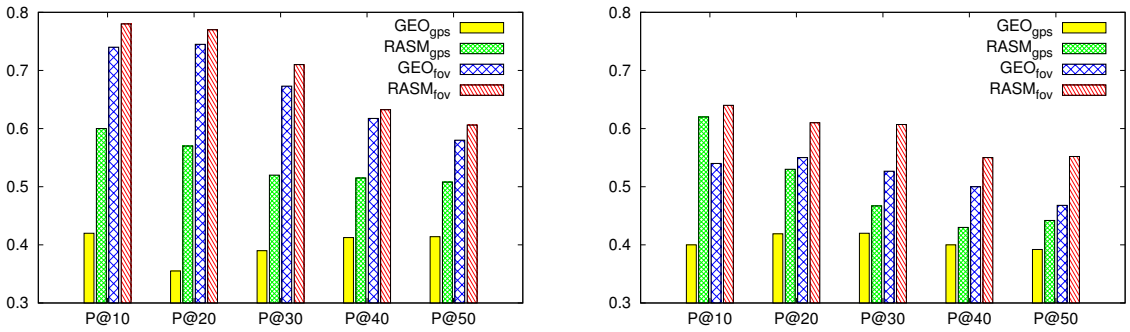
As shown in Table IV, the MAP decreased when we replaced one component by an existing one. This demonstrates the effectiveness and the indispensability of our proposed approach.

4) System Efficiency: We performed the retrieval experiments on a desktop computer with a 3.20 GHz dual core CPU and 4 GB of main memory. The comparison of the execution time for feature extraction is reported in Table V. For each query that we executed, we recorded the retrieval latency which includes the similarity calculation and the result ranking. The average value is reported in Table VI.

TABLE V: The comparison of the execution time for feature extraction per image.

Feature	*GEO*	*Color*	*Texture*	*Gist*	*OB*	*BoS*
Time	0.01 ms	0.06 s	0.12 s	0.46 s	4.68 s	4.682 s

In comparison of the execution time for feature extraction, the encoding of the proposed geo-features is highly efficient



(a) Query-by-video. (b) Query-by-image.

Fig. 9: P@n comparison based on different availability of geo-metadata.

TABLE VI: The comparison of the average retrieval latency.

Method	*GEO*	*CRLF*	*CRGV*	*OB*	*BoS*	*RASM*
By-video	6 ms	512 ms	525 ms	927 ms	11 ms	295 ms
By-image	6 ms	83 ms	98 ms	185 ms	11 ms	64 ms

as the calculation is only based on the camera location and orientation. In contrast, the time complexity for visual feature extraction is much higher. As can be seen, the low-level visual features such as color and texture would cost dozens of milliseconds for extraction, while the more descriptive ObjectBank representation would cost more than four seconds. *BoS* cost slightly more than *OB* as the former takes an extra step by soft encoding each frame to its nearest neighbors in the dictionary. Our proposed model can achieve high MAP while maintaining good efficiency. With the help of the proposed content-oriented geo-feature, effective retrieval performances can be achieved by using only the less descriptive low-level visual features, and thus the time complexity is greatly reduced.

As aforementioned, the videos in our system are indexed using inverted files based on the geographic regions. Therefore, only the geo-relevant videos are processed for similarity calculations. Method *GEO* is highly efficient because the high-dimensional visual features are not utilized in the similarity calculations. The visual approach BoS reduced the time complexity by generating a visual descriptor per video segment instead of per frame. Method *OB* is the least efficient due to its high dimensionality compared with color, texture, and Gist used in other approaches. For hybrid approaches, the visual feature comparison is always the major cost for both storage and computation. Let $\bar{n}$ denote the average number of keyframes in a video, then the complexity for the visual similarity calculation in *CRLF*, *CRGV*, and the late fusion approach will be $\mathrm{O}\left(\bar{n}^2\right)$. Different from the above methods where a pairwise comparison between keyframes is required, our proposed approach *RASM* reduces the computational costs by geographic indexing where only the local visual similarities of each region are computed. If the keyframes of a video are divided into an average of $\bar{k}$ region groups, the time complexity will be reduced to $\mathrm{O}\left(\bar{n}^2/\bar{k}\right)$. Comparatively, most of the previous work focused on the compact video representations that support efficient visual indexing [24], [31]. It is worth emphasizing that such techniques are parallel to our model, which can be integrated on the region-level after frames are geographically indexed. The geographic and the visual indexing complement with each other in a large video database. Considering the limitations on the availability of current geo-referenced videos, discussions of integrating efficient approximate visual similarity measure are left as part of the future work.

D. Evaluation on Semantic Annotation

The geotags generated based on OSM were compared to the enriched tags generated by the approaches, $Enriched_A$ and $Enriched_B$, introduced in Section V. Images were filtered based on the geo-locations. Therefore, annotation accuracy can be further improved by geographic restrictions.

Due to the lack of enough sensor data, it is difficult to use acknowledged datasets (*e.g.*, TRECVID, http://trecvid.nist.gov/) for our experiments. Additionally, the geo-referenced videos are usually organized by the geographic information without any ground-truth textual labels on a list of concepts. Therefore, we carried out a user study to evaluate the quality of the generated tags. Ten video clips from different regions were selected. Without loss of generality, only the top ten tags generated using different methods were preserved. 26 volunteers who are familiar with the regions where the videos were taken participated in this user study. They were requested to watch the video carefully and then give scores based on the (1) relevance and (2) diversity of the generated tags (1-least, 10-most). The results of this user study are presented in Fig. 11 and 12. As can be seen, while the three methods achieved comparative results in terms of relevance, the diversity of the tags has been greatly improved by applying tag refinement techniques.

To demonstrate the annotation results, we show an example in Fig. 10. The first row lists the tags generated by Shen *et al.*'s approach. It relies on the OSM which may have uneven building and detail coverage. Therefore, the annotation errors can be caused by missed obstacles and occlusions. Despite the issues above, its overall precision is high. One limitation of this method is that the type of tags it selects is uniform, *i.e.*, the name of the surrounding places and buildings, which is usually

OSM	The Fullerton Hotel, Maybank Tower, One Fullerton, HSBC Building, Hitachi Tower, Tung Centre, Marina Bay Reservoir, North Tower, One Raffles Quay, Change Alley
Enriched$_A$	The Fullerton Hotel, Singapore, One Fullerton, Cityscape, Street, Buildings, City, Water, Asia, Urban
Enriched$_B$	The Fullerton Hotel, Boat Quay, Artwork, Statues, Cavenagh Bridge, Singapore River, One Fullerton, HSBC, Maybank, Esplanade Bridge

Fig. 10: Illustration of the top ten tags generated for an example video.

not the most preferred situation by users. Comparatively, the enriched tags listed in the second and third rows are more consistent with the user preferences. Not only the semantics have been greatly enriched, the wrongly assigned geotags due to occlusions have also been removed. As we can see, some of the tags (*e.g.*, *Asia*) generated by $Enriched_A$ may be too general for a specific video clip. Comparatively, $Enriched_B$ tends to favor more specific tags by learning the tag relevance from its visual neighbors. This is also consistent with the results of the user study. The participants generally agreed that $Enriched_B$ performed the best in terms of tag diversity. However, $Enriched_B$ may suffer from a slight decrease in precision compared with $Enriched_A$ as shown in Fig. 11.

VII. Conclusions and Future Work

This paper proposed the generation of content-oriented geo-features to facilitate video annotation and search. It does not focus on one specific visual similarity measure, rather it shows that the innovative fusion of visual and geo features provides improved performance over the existing approaches. A novel hybrid model is proposed as video representation, describing both the video geographic coverage and the region-aware representative visual features. Additionally, we propose to construct a geo-codebook by utilizing the information available from the geo-information services to segment an area into a set of coherent regions. It overcomes the limitations of a grid-based codebook, based on which the geographic coverage of a video can be better encoded. Lastly, we developed a video search prototype based on our proposed hybrid model. To evaluate its performance, we compared it to existing approaches and a user study was carried out accordingly. The good results demonstrate the effectiveness of our proposed approaches. Toward a more effective video retrieval system, more efforts will be made on the correction of geographic metadata and the acceleration of visual similarity calculations.

Acknowledgements

This research was supported in part by the National Natural Science Foundation of China under Grant no. 61472266 and by the National University of Singapore (Suzhou) Research Institute, 377 Lin Quan Street, Suzhou Industrial Park, Jiang Su, People's Republic of China, 215123.

References

[1] "Youtube press room, statistics," http://www.youtube.com/yt/press/statistics.html, March 2014.
[2] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua, "Bayesian Video Search Reranking," in *ACM Multimedia*, 2008, pp. 131–140.
[3] J. Liu, W. Lai, X.-S. Hua, Y. Huang, and S. Li, "Video Search Re-Ranking via Multi-Graph Propagation," in *ACM Multimedia*, 2007, pp. 208–217.
[4] R. Jain and P. Sinha, "Content Without Context is Meaningless," in *ACM Multimedia*, 2010, pp. 1259–1268.
[5] C. Kumar, "Relevance and Ranking in Geographic Information Retrieval," in *BCS-IRSG FDIA*, 2011, pp. 2–7.
[6] L. S. Kennedy and M. Naaman, "Generating Diverse and Representative Image Search Results for Landmarks," in *WWW*, 2008, pp. 297–306.
[7] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the World's Photos," in *WWW*, 2009, pp. 761–770.
[8] J. Kleban, E. Moxley, J. Xu, and B. S. Manjunath, "Global Annotation on Georeferenced Photographs," in *ACM Image and Video Retrieval*, 2009, pp. 1–8.
[9] J. Kamahara, T. Nagamatsu, and N. Tanaka, "Conjunctive Ranking Function using Geographic Distance and Image Distance for Geotagged Image Retrieval," in *ACM GeoMM*, 2012, pp. 9–14.
[10] S. Arslan Ay, R. Zimmermann, and S. H. Kim, "Viewable Scene Modeling for Geospatial Video Search," in *ACM Multimedia*, 2008, pp. 309–318.
[11] S. Arslan Ay, R. Zimmermann, and S. Kim, "Relevance Ranking in Georeferenced Video Search," *Multimedia Systems*, vol. 16, pp. 105–125, 2010.
[12] Y. Kim, J. Kim, and H. Yu, "GeoSearch: Georeferenced Video Retrieval System," in *ACM SIGKDD*, 2012, pp. 1540–1543.
[13] A. G. Hauptmann and M. G. Christel, "Successful Approaches in the TREC Video Retrieval Evaluations," in *ACM Multimedia*, 2004, pp. 668–675.
[14] M. Campbell, A. Haubold, S. Ebadollahi, D. Joshi, M. R. Naphade, A. P. Natsev, J. Seidl, J. R. Smith, K. Scheinberg, J. Tešic, and L. Xie, "IBM Research TRECVID-2006 Video Retrieval System," 2006.
[15] O. A. B. Penatti, L. T. Li, J. Almeida, and R. da S. Torres, "A Visual Approach for Video Geocoding Using Bag-of-Scenes," in *ACM ICMR*, 2012, pp. 53:1–53:8.
[16] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei, "Object Bank: An Object-Level Image Representation for High-Level Visual Recognition," *International Journal of Computer Vision*, pp. 20–39, 2014.
[17] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video Search Reranking through Random Walk over Document-Level Context Graph," in *ACM Multimedia*, 2007, pp. 971–980.
[18] Y. Yin, Z. Shen, L. Zhang, and R. Zimmermann, "Spatial-Temporal Tag Mining for Automatic Geospatial Video Annotation," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 11, pp. 29:1–29:21, 2015.
[19] Y. Yin, B. Seo, and R. Zimmermann, "Content vs. Context: Visual and Geographic Information Use in Video Landmark Retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 11, no. 3, pp. 39:1–39:21, 2015.
[20] S. Liao, X. Li, X. Wang, and X. Du, "Building Geo-aware Tag Features for Image Classification," in *IEEE ICME*, 2014, pp. 1–6.
[21] B. Zhang, Q. Li, H. Chao, B. Chen, E. Ofek, and Y.-Q. Xu, "Annotating and Navigating Tourist Videos," in *ACM SIGSPATIAL GIS*, 2010, pp. 260–269.
[22] Z. Shen, S. Arslan Ay, S. H. Kim, and R. Zimmermann, "Automatic Tag Generation and Ranking for Sensor-Rich Outdoor Videos," in *ACM Multimedia*, 2011, pp. 93–102.
[23] S. Cheung and A. Zakhor, "Efficient Video Similarity Measurement and Search," in *Image Processing*, 2000, pp. 85–88.
[24] H. T. Shen, B. C. Ooi, and X. Zhou, "Towards Effective Indexing for Very Large Video Sequence Database," in *ACM SIGMOD*, 2005, pp. 730–741.
[25] J. Yuan, Z.-J. Zha, Y.-T. Zheng, M. Wang, X. Zhou, and T.-S. Chua, "Learning Concept Bundles for Video Search with Complex Queries," in *ACM Multimedia*, 2011, pp. 453–462.
[26] L. Cao, Z. Li, Y. Mu, and S.-F. Chang, "Submodular Video Hashing: A Unified Framework Towards Video Pooling and Indexing," in *ACM Multimedia*, 2012, pp. 299–308.
[27] X. Li, C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Fusing Concept Detection and Geo Context for Visual Search," in *ACM ICMR*, 2012, pp. 4:1–4:8.

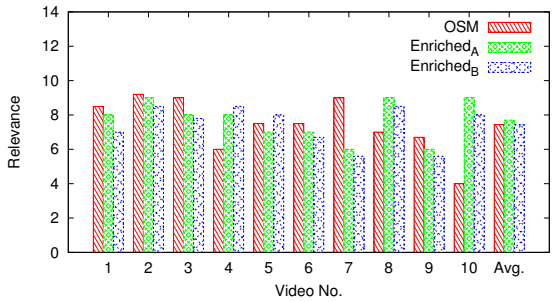


Fig. 11: Relevance comparison of the initial OSM and the enriched Flickr tags.

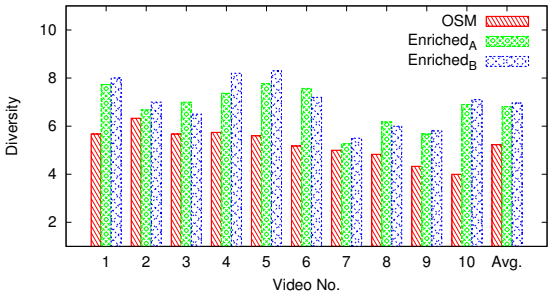


Fig. 12: Diversity comparison of the initial OSM and the enriched Flickr tags.

[28] X.-Y. Wei and C.-W. Ngo, "Ontology-enriched Semantic Space for Video Search," in *ACM Multimedia*, 2007, pp. 981–990.
[29] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to Predict Where Humans Look," in *Computer Vision*, 2009, pp. 2106–2113.
[30] R. H. Van Leuken, L. Garcia, X. Olivares, and R. van Zwol, "Visual Diversification of Image Search Results," in *WWW*, 2009, pp. 341–350.
[31] S. Cheung and A. Zakhor, "Efficient Video Similarity Measurement with Video Signature," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, pp. 59–74, 2003.
[32] B. Thomee and A. Rae, "Uncovering Locally Characterizing Regions Within Geotagged Data," in *WWW*, 2013, pp. 1285–1296.
[33] S. Intagorn and K. Lerman, "Learning Boundaries of Vague Places from Noisy Annotations," in *ACM SIGSPATIAL GIS*, 2011, pp. 425–428.
[34] A. Ballatore, M. Bertolotto, and D. Wilson, "Geographic Knowledge Extraction and Semantic Similarity in OpenStreetMap," *Knowledge and Information Systems*, pp. 61–81, 2013.
[35] M. E. Newman, "Analysis of Weighted Networks," *Physical Review E*, p. 056131, 2004.
[36] T. H. Silva, P. O. S. Vaz de Melo, J. M. Almeida, J. Salles, and A. A. F. Loureiro, "A Comparison of Foursquare and Instagram to the Study of City Dynamics and Urban Social Behavior," in *ACM SIGKDD UrbComp*, 2013, pp. 1–8.
[37] M. De Choudhury, M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, and C. Yu, "Automatic Construction of Travel Itineraries Using Social Breadcrumbs," in *ACM Conference on Hypertext and Hypermedia*, 2010, pp. 35–44.
[38] H. J. Seo and P. Milanfar, "Static and Space-time Visual Saliency Detection by Self-Resemblance," *Journal of Vision*, vol. 9, 2009.
[39] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global Contrast based Salient Region Detection," *IEEE Transections on Pattern Analysis and Machine Intelligence*, vol. 37, pp. 569–582, 2015.
[40] M. Wang, X.-S. Hua, J. Tang, and R. Hong, "Beyond Distance Measurement: Constructing Neighborhood Similarity for Video Annotation," *IEEE Transactions on Multimedia*, vol. 11, pp. 465–476, 2009.
[41] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with Local and Global Consistency," *NIPS*, vol. 16, pp. 321–328, 2004.
[42] L. Ballan, M. Bertini, A. Del Bimbo, M. Meoni, and G. Serra, "Tag Suggestion and Localization in User-generated Videos Based on Social Knowledge," in *ACM SIGMM Workshop on Social Media*, 2010, pp. 3–8.
[43] X. Li, C. Snoek, and M. Worring, "Learning Social Tag Relevance by Neighbor Voting," *IEEE Transactions on Multimedia*, vol. 11, pp. 1310–1322, 2009.
[44] L. Ballan, M. Bertini, T. Uricchio, and A. Del Bimbo, "Social Media Annotation," in *International Workshop on Content-Based Multimedia Indexing*, 2013, pp. 229–235.
[45] B. Manjunath and W. Ma, "Texture Features for Browsing and Retrieval of Image Data," *IEEE Transections on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 837–842, 1996.
[46] M. Stricker and M. Orengo, "Similarity of Color Images," in *SPIE Storage and Retrieval for Image and Video Databases III*, 1995, pp. 381–392.
[47] A. Oliva and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," *Int. J. Comput. Vision*, vol. 42, pp. 145–175, 2001.
[48] N. O'Hare, C. Gurrin, G. Jones, and A. Smeaton, "Combination of Content Analysis and Context Features for Digital Photograph Retrieval," in *European Workshop on Integration of Knowledge, Semantics and Digital Media Technology*, 2005, pp. 323–328.
[49] C. Hauff, "A Study on the Accuracy of Flickr's Geotag Data," in *ACM SIGIR*, 2013, pp. 1037–1040.

Yifang Yin received the B.E. degree from the Department of Computer Science and Technology, Northeastern University, Shenyang, China, in 2011. She is currently working towards the PhD. degree from the School of Computing, National University of Singapore, Singapore. She worked as a Research Intern at the Incubation Center, Research and Technology Group, Fuji Xerox Co., Ltd., Japan, from October, 2014 to March, 2015. Her research interests include geotagged video annotation and retrieval, geo-metadata correction and video summarization.

Yi Yu is currently an assistant professor with National Institute of Informatics (NII), Japan. Before joining NII, she was a senior research fellow with School of Computing, National University of Singapore. Her research covers large-scale multimedia information processing and analysis, location-based mobile media service and social media analysis. Yu received a Ph.D. in Information and Computer Science from Nara Women's University, Japan.

Roger Zimmermann (S'93M'99SM'07) received the M.S. and Ph.D. degrees from the University of Southern California (USC) in 1994 and 1998. He is currently an associate professor in the Department of Computer Science at the National University of Singapore (NUS). He is also a deputy director with the Interactive and Digital Media Institute (IDMI) at NUS and a co-director of the Centre of Social Media Innovations for Communities (COSMIC). His research interests are in the areas of streaming media architectures, distributed and peer-to-peer systems, mobile and geo-referenced video management, collaborative environments, spatio-temporal information management, and mobile location-based services. He has coauthored a book, six patents, and more than 150 conference publications, journal articles, and book chapters. He is a member of ACM.