Content vs. Context: Visual and Geographic Information Use in Video Landmark Retrieval

YIFANG YIN, National University of Singapore
BEOMJOO SEO, Hongik University
ROGER ZIMMERMANN, National University of Singapore

Due to the ubiquity of sensor-equipped smartphones, it has become increasingly feasible for users to capture videos together with associated geographic metadata, for example the location and the orientation of the camera. Such contextual information creates new opportunities for the organization and retrieval of geo-referenced videos. In this study we explore the task of landmark retrieval through the analysis of two types of state-of-the-art techniques, namely *media-content-based* and *geocontext-based* retrievals. For the content-based method, we choose the *Spatial Pyramid Matching* (SPM) approach combined with two advanced coding methods: *Sparse Coding* (SC) and *Locality-Constrained Linear Coding* (LLC). For the geo-based method, we present the *Geo Landmark Visibility Determination* (GeoLVD) approach which computes the visibility of a landmark based on intersections of a camera's *field-of-view* (FOV) and the landmark's geometric information available from *Geographic Information Systems* (GIS) and services. We first compare the retrieval results of the two methods, and discuss the strengths and weaknesses of each approach in terms of precision, recall and execution time. Next we analyze the factors that affect the effectiveness for the content-based and the geo-based methods, respectively. Finally we propose a hybrid retrieval method based on the integration of the visual (content) and geographic (context) information, which is shown to achieve significant improvements in our experiments. We believe that the results and observations in this work will enlighten the design of future geo-referenced video retrieval systems, improve our understanding of selecting the most appropriate visual features for indexing and searching, and help in selecting between the most suitable methods for retrieval based on different conditions.

Categories and Subject Descriptors: H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*

General Terms: Experimentation

Additional Key Words and Phrases: Content-based analysis, geo-referenced videos, landmark retrieval

ACM Reference Format:
Yifang Yin, Beomjoo Seo, and Roger Zimmermann. 2015. Content vs. context: Visual and geographic information use in video landmark retrieval. ACM Trans. Multimedia Comput. Commun. Appl. 11, 3, Article 39 (January 2015), 21 pages.
DOI: http://dx.doi.org/10.1145/2700287

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). This work was also supported by the Hongik University new faculty research support fund.
Authors' addresses: Y. Yin, NUS School of Computing, Computing 1, 13 Computing Drive, Singapore 117417; email: yifang@comp.nus.edu.sg; B. Seo; email: bseo@hongik.ac.kr; R. Zimmermann; email: rogerz@comp.nus.edu.sg.

DOI: http://dx.doi.org/10.1145/2700287

1. INTRODUCTION

The retrieval of landmark sequences from video collections still remains a very challenging task. In traditional video repositories, landmark retrieval is often conducted by matching query keywords to manually entered text associated with videos such as titles, tags and descriptions. However, it is challenging for users to obtain satisfactory search results via keywords because textual annotations can be sometimes ambiguous, sparse and noisy [Huang et al. 2010; Tian et al. 2008]. Content-based visual information retrieval offers a promising approach for landmark search. In recent years the *bag-of-words* (BoW) model [Csurka et al. 2004], which was inspired by the success of text-based retrieval, has been extremely popular in a variety of visual retrieval and categorization tasks. The original *BoW* approach has subsequently been improved in its performance in various ways [Lazebnik et al. 2006; Yang et al. 2009; Wang et al. 2010]. Such content-based retrieval, however, has one drawback that hinders scalability, namely high computational complexity due to extensive signal-level processing. Moreover, it is susceptible to environmental conditions associated with an image, for example its illumination and camera recording angle [Kuo et al. 2012; Jain and Sinha 2010].

Since content-based retrieval is sometimes struggling to achieve satisfactory results, researchers have begun to utilize *contextual information* as an alternative or supplement to visual information. For outdoor videos and images, geographic information is especially useful. When performing a landmark retrieval task among a set of GPS-tagged images, geoclustering is usually applied at an early stage [Kennedy and Naaman 2008; Avrithis et al. 2010]. Such an approach helps not only in the organization of large image collections, but also in achieving better efficiency and accuracy. With today's sensor-equipped smartphones it is also possible to tag recorded videos with a fine-grained, continuous stream of extended geographic properties that relate to the captured camera scenes [Arslan Ay et al. 2008]. With this method the viewable scene is modeled based on the sequence of location and orientation information collected from GPS and compass sensors. This metadata is then utilized for auto-tagging and searching large collections of community-generated videos [Arslan Ay et al. 2010; Shen et al. 2011]. The principles of the geo-based technique can also be applied to landmark retrieval. One challenge is that its performance is influenced by the accuracy of the sensor data.

In this study, we evaluate, compare, and finally integrate two major types of landmark retrieval techniques: (a) the content-based and (b) the geo-based-approaches. Note that we do not utilize textual metadata as it differs from visual content and geocontext in terms of granularity, that is, textual annotations such as titles and tags are usually not localized to frames. Moreover, textual annotations can be ambiguous and noisy and hence their accuracy is difficult to assess. The video collection we use in our experiments is *geo-referenced*, meaning the location and orientation of the cameras were recorded and associated with the video streams as metadata that can subsequently be used for geo-based retrieval. For content-based retrieval, videos need to be preprocessed such that the visual feature information is extracted, coded, and stored.

We first compare two state-of-the-art content-based methods, namely, *Spatial Pyramid Matching with Sparse Coding* (ScSPM) [Yang et al. 2009] and *Locality-Constrained Linear Coding* (LLC) [Wang et al. 2010], with a geo-based method which we refer to as *Geo Landmark Visibility Determination* (GeoLVD), in terms of precision, recall, and execution time, respectively. We selected these methods as representatives because they currently exemplify the state-of-the-art and are superior in their own fields. Both *ScSPM* and *LLC* enrich the traditional *BoW* with spatial information and the advanced coding techniques they adopt not only accelerate the processing speed but also significantly improve the effectiveness. *GeoLVD* computes the visibility of a landmark based on intersections of a camera's field-of-view and the landmark's geometric information

available from GIS. It utilizes the state-of-the-art *FOVScene* [Arslan Ay et al. 2008] to model the camera's field-of-view and its effectiveness is well supported by experimental results. Second, we analyze the detailed factors that can affect the retrieval effectiveness. For the content-based method, we investigate the influence of selecting a representative training set and the impact of the diversity of the video frames. We also seek better sources for training images and propose to use Google StreetView as a supplement to Flickr, a combination which is shown to be effective in our experiments. For the geo-based *GeoLVD* method we analyze the influence of the accuracy of a video's geographic metadata and the level of detail of the information we can obtain from geographic information system sources. Finally, we propose a hybrid retrieval method based on the integration of visual and geographic information. Experiments show that such a combination is compelling and achieves the best performance in the landmark retrieval task.

The contributions of this study are summarized in the following three aspects.

(i) We compare state-of-the-art retrieval techniques in terms of precision, recall and execution time, respectively, analyze the strength and weakness of each method, and discuss how to select the most suitable retrieval method according to video conditions and system requirements.
(ii) We investigate the factors that affect the retrieval effectiveness, measure and compare their influence through experiments, and propose methods to reduce their adverse influence when it is possible.
(iii) We propose a hybrid retrieval method by integrating visual and geographic information. Experiments show that the integration achieves significant improvements in terms of precision and recall.

The rest of the article is organized as follows. We first report on important related work in Section 2, introduce details of the evaluated methods in Section 3, and describe the dataset and experimental settings in Section 4. The retrieval results are reported and discussed in Section 5. Next we analyze the factors that influence the retrieval effectiveness in Sections 6 and 7 for the content-based and geo-based methods, respectively. A hybrid retrieval method is proposed in Section 8 by integrating the visual and geographic information. Finally, we draw conclusions and report on future research directions in Section 9.

2. RELATED WORK

Geotagging is increasingly popular for media such as images and videos posted on websites and blogs [Luo et al. 2011]. Geographical location tags help users to localise videos, allowing the media to be anchored to real world locations [Rae et al. 2011]. The *MediaEval Placing Task* is held annually for participants to attempt to automatically assign latitude and longitude coordinates to each of the provided test videos [Kelm et al. 2011]. Nowadays, with an increasing number of devices being available that can automatically encode geotags, it has become easier and more efficient to record the geo-metadata at the time when videos are taken. A variety of methods and solutions benefit from the presence of geographically relevant metadata. Liu et al. [2005] presented a sensor enhanced video annotation system (referred to as *SEVA*) which enables searching videos for the presence of particular objects. However this approach requires a controlled environment where a sensor is attached to every object. Simon et al. [2007] presented an application framework that retrieves the visible objects within the user's viewable scene in the real world. Arslan Ay et al. [2008] proposed a viewable scene model, based on camera location and orientation, to describe the viewable region within a video. This viewable scene model was further extended for efficient tagging and searching in other work [Shen et al. 2011; Arslan Ay et al. 2010; Kim et al. 2012]. One challenge is the

occurance of GPS and compass errors and therefore techniques based on geographic information may sometimes face performance issues. Zhang et al. [2010] proposed an annotation and navigation system for tourist videos based on video tracks and orientation. The method can calibrate, or even obtain, position and orientation information by registering videos to geo-referenced 3D models. It brought awareness to the importance of geographic metadata, especially for tourist videos.

In the computer vision domain, the bag-of-words method is the current state-of-the-art approach for landmark image retrieval [Csurka et al. 2004]. The most popular choice for feature extraction in the *BoW* model is the *Scale-Invariant Feature Transform* (SIFT) descriptor [Lowe 2004]. It has been reported that, in terms of landmark recognition, *SIFT* outperforms not only global features such as color and texture, but also other local features such as *Speeded Up Robust Features* (SURF) and *Multi-Scale Oriented Patches* (MSOP) [Amato et al. 2012; Yap et al. 2010]. Yap et al. [2010] also showed that *dense-SIFT* works better than *sparse-SIFT*, and an enhanced *BoW* integrated with multiresolution patches and *dense-SIFT* achieves the best performance.

As the traditional *BoW* approach discards the spatial information of local descriptors, the descriptive power of its image representation is severely limited. Subsequently, efforts have been made to encode the spatial information into image content descriptions [Hoàng et al. 2010; Penatti et al. 2014]. Lazebnik et al. [2006] proposed a spatial pyramid matching technique for natural scene categorization. Advanced coding techniques have also been proposed, which better encode the original feature descriptors based on the vocabulary basis to yield significant performance improvements [Yang et al. 2009; Wang et al. 2010]. Endeavors to enrich the *BoW* model with spatial information from other perspectives have been tried as well, such as using homography mappings that geometrically connect pairs of images [Chum et al. 2007; Gavves et al. 2012]. The idea of expanding 2D images into 3D landmark models for the task of landmark recognition has also been studied [Hao et al. 2012]. However, performance improvements are achieved by adopting more complex spatial models with a larger vocabulary, at the expense of high memory and computational costs.

In the last several years, an important trend has emerged within the multimedia and computer vision communities in an increasing emphasis on modeling and use of contextual information. Researchers began to utilize geographic data as supplement to visual information. Several methods have recently been proposed for landmark recognition and retrieval that integrate geographic information with content analysis, but with different goals. Zheng et al. [2009] presented a web-scale landmark recognition engine that organizes, models and recognizes landmarks on the scale of the entire planet Earth. Avrithis et al. [2010] proposed a system that can retrieve not only landmarks but also nonlandmark images in collections of community photos by constructing a 2D scene map for each view cluster and preserving details from all the reference images while discarding repeated visual features. Chen et al. [2011a] addressed the problem of city-scale landmark recognition from cell phone images. More advanced content and context integration techniques for mobile landmark recognition have been proposed to achieve better performance as well [Chen et al. 2011b; Li and Yap 2012].

Most of the recent approaches on landmark recognition and retrieval focus either on the landmark organization and modeling from large community photo collections, or on the real-time landmark recognition within an image taken from a mobile phone. Our work addresses a different aspect of landmark retrieval from geo-referenced video collections. It also differs from previous video retrieval techniques [Souvannavong et al. 2005; Feng and Manmatha 2008] in that it has a more specific focus on landmark retrieval and proposes landmark recognition techniques suitable for videos, not just images. Recently, Penatti et al. [2012] proposed a novel video representation model, called Bag-of-Scenes, which uses scenes as the basic elements to represent a video.

The method has shown promising results in video geocoding, but its performance in video retrieval still remains unknown. Moreover, the dictionary of scenes is predefined, so issues may arise when retrieving relevant segments of an arbitrary landmark that differs from any of the scenes in the dictionary.

3. LANDMARK RETRIEVAL METHODS

The initial step for video landmark retrieval is to determine the landmark's visibility in any given frame. We describe the algorithmic fundamentals for two different retrieval paradigms to recognize landmarks from a collection of community contributed videos. Section 3.1 first describes two existing state-of-the-art content-based methods, and then Section 3.2 introduces a context-based method that utilizes geographic properties.

3.1. Landmark Retrieval from Visual Cues and Features

Recently the *BoW* model [Csurka et al. 2004] has been extremely popular for use in a variety of visual retrieval and categorization tasks because of its high classification quality. The method treats an image as a collection of orderless appearance descriptors extracted from local patches, quantizes them into discrete visual words, and then computes a compact histogram representation for semantic image classification. Subsequently *SVM* classifiers are constructed from the labelled *BoW* representations. Here we model the landmark retrieval task as a two-class (positive vs. negative) classification problem. In a retrieval session, our system selects the *SVM* trained for the target landmark, and then scans and ranks the frames based on the probability scores output by the selected *SVM*.

One limitation, however, is that the *BoW* approach discards the spatial information of local descriptors, which severely limits the descriptive power of the image representation. Therefore the *SPM* technique has been proposed to overcome this issue by Lazebnik et al. [2006] for natural scene categorization. The approach creates a spatial pyramid representation by partitioning the image into increasingly fine subregions and computing histograms of local features found inside each subregion. *SPM* shows significantly improved performance on challenging scene categorization tasks. However, researchers have empirically found that it only works well with a particular type of nonlinear Mercer kernels, for instance, the *Chi-square* kernel [Yang et al. 2009]. As a consequence, the nonlinear classifier imposes additional computational complexity, namely $\mathcal{O}(n^3)$ in training and $\mathcal{O}(n)$ in testing, where n is the number of support vectors. New coding techniques have been proposed to make it work well with simple linear *SVMs*, which can dramatically improve the scalability of the training phase and the speed of testing. Two linear *SPM*s with advanced coding techniques are tested in our experiments. The algorithmic details are described in Sections 3.1.1 and 3.1.2.

3.1.1. Linear Sparse Coding. A method called *ScSPM* [Yang et al. 2009] has been proposed to compute the spatial-pyramid image representation based on *sparse coding* (SC), instead of the *K*-means *vector quantization* (VQ) in the traditional *SPM*. The approach is naturally derived by relaxing the restrictive cardinality constraint of *VQ*. Furthermore, it uses max pooling, which is more robust to local spatial translations and more biologically plausible, rather than the average pooling adopted in the original *SPM*.

Let X be a set of D-dimensional local descriptors extracted from an image, that is, $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{D\times N}$, and let $B = [b_1, b_2, \ldots, b_M] \in \mathbb{R}^{D\times M}$ be a codebook with M entries. The *SC* in *ScSPM* can be described as solving the following problem:

$$\underset{C}{\operatorname{argmin}} = \sum_{i=i}^{N} \parallel x_i - Bc_i \parallel^2 + \lambda \parallel c_i \parallel_l, \tag{1}$$

where $C = [c_1, c_2, \ldots, c_N] \in \mathbb{R}^{M \times N}$ is the set of sparse codings for X. The restrictive cardinality constraint of VQ is relaxed by using a sparsity regularization term, which is the l norm of c_i in this case. The codebook training is equivalent to finding the optimization of problem Equation (1), which can be solved by algorithms such as the *feature-sign search* algorithm.

It has been found that *ScSPM* works well with simple linear *SVMs*, which means it remarkably reduces the complexity of *SVMs* to $\mathcal{O}(n)$ in training and a constant in testing, and even improves the classification accuracy. However, the coding speed is relatively slow.

3.1.2. Locality-Constrained Linear Coding. Wang et al. [2010] proposed an approach that enables *SPM* to work with locality-constrained linear coding, referred to as *LLC*. This approach also adopts max pooling, and it utilizes the locality constraints to project each descriptor into its local-coordinate system. The *LLC* code uses the following criteria:

$$\begin{aligned} &\underset{C}{\operatorname{argmin}} = \sum_{i=i}^{N} \parallel x_i - Bc_i \parallel^2 + \lambda \parallel d_i \odot c_i \parallel^2 \\ &s.t. \quad 1^{\mathrm{T}} c_i = 1, \quad \forall i, \end{aligned} \tag{2}$$

where $\odot$ denotes element-wise multiplication, and $d_i \in \mathbb{R}^M$ is the locality adaptor that allocates different freedom for each basis vector proportional to its similarity to the input descriptor x_i.

Just like with *ScSPM*, the codebook can be trained using *LLC* coding criteria shown in Equation (2). In effect, it can also be trained using a simple K-means clustering method, since experiments have shown that the codebook generated by K-means clustering can produce satisfactory accuracy. More extensive experiments on the selection of codebooks were carried out by Viitaniemi and Laaksonen [2008]. A fast approximated *LLC* method has been proposed as well to further speed up the encoding process. This efficiency significantly adds to the practical value of *LLC* for real-world applications.

3.2. Landmark Retrieval from Geographic Information

Due to technological advances, sensor-equipped smart phones have made it easy to record geo-referenced videos through the use of popular apps. The location and direction of the camera can be obtained from the built-in GPS and compass sensors of a smartphone. Furthermore, geographic information about landmarks can be retrieved from map sources such as OpenStreetMap[1], which is a free service that provides editable maps of the world. Next, we will introduce the details of a landmark retrieval technique based on geographic information.

3.2.1. Viewable Scene Model. Presented with the geographic information associated with a video frame, here we utilize the viewable scene model (referred to as *FOVScene*) proposed by Arslan Ay et al. [2008] to describe the visible scene based on a camera's *field-of-view* (FOV). The 3-dimensional $FOVScene(P, \vec{d}, \theta, \phi, R)$ model is illustrated in Figure 1, with the following parameters: (1) the camera position P, (2) the camera direction (i.e., compass) vector $\vec{d}$, (3) the horizontal and vertical camera viewable angles θ and ϕ which describe the angular extent of the scene filmed by the camera, and (4) the far visible distance R which is the maximum distance at which a large object within the camera's field-of-view can be recognized. In this study, the position (latitude/longitude) and orientation information of the camera are associated with video streams

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

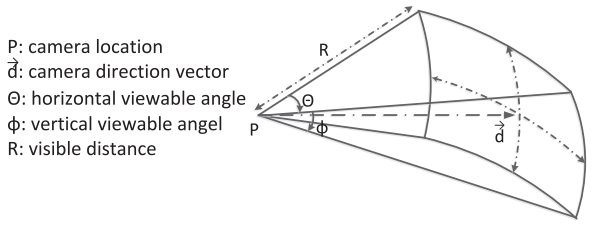


Fig. 1. Illustration of the 3D *FOVScene* model.

at a fine-granular level as metadata, that is, at each—or every few—frames. For simplicity, we assume that the camera is always level, that is, the direction vector $\vec{d}$ only stays on the horizontal plane. The parameters θ, ϕ and R are constants that can be estimated from the optics of the camera used for video recording [Hecht 2001].

3.2.2. Determination of Landmark Visibility. Given the name of a landmark and a video frame, the task here is to determine the landmark's visibility in the frame. The geometry of the viewable scene of the frame G_{frame} is modeled as $FOVScene(P, \vec{d}, \theta, \phi, R)$. The geometry of the landmark $G_{landmark}$ can be retrieved from, for example, OpenStreetMap. We also need to consider the situation where sometimes the existence of other buildings may hide the landmark from the camera's sight, and it is therefore necessary to retrieve the geometry of all relevant objects within the same region as the landmark for occlusion checks. Let $\{o_1, o_2, \ldots, o_k\}$ be the set of all relevant objects, then $\{G_{o_1}, G_{o_2}, \ldots, G_{o_k}\}$ represents their corresponding geometry set.

Given this geometry information, we introduce an algorithm termed *GeoLVD* to determine the visibility of a certain landmark within a frame. This algorithm is inspired by the 3D visibility query processor proposed by Shen et al. [2011], but differs mainly in the following three aspect.

—The *GeoLVD* algorithm determines the visibility of a given landmark, while Shen's method aims to find all the visible objects within the viewable scene.
—The *GeoLVD* algorithm considers the landmark to be either visible or invisible, while Shen et al. further classify the visible objects into two subcategories: front objects and vertically visible objects.
—The *GeoLVD* algorithm uses an *R*-tree* index structure to organize the geometry of all the relevant objects, while Shen's method does not mention any advanced data structure adopted for spatial indexing.

Instead of assigning scores to frames based on the ratio of viewable angles of the target landmark, we define the output of *GeoLVD* to be binary, that is, the output is either zero or one, where zero (one) indicates that the landmark is invisible (visible). Our rationale for using a binary output is that in our experience the effectiveness of the geo-based method highly depends on the accuracy of the geographic metadata. We found that the use of more elaborate geometry calculations provides no significant benefit for the correction of the variations induced by noise which the geographic metadata intrinsically possesses. Figure 2 shows an example of the *GeoLVD* method in (a) Google Earth together with its corresponding (b) projection on the 2D plane. Assume that the landmark to be retrieved is the Marina Bay Sands hotel. The main steps of the algorithm are as follows. First it computes the field-of-view towards the landmark within the *FOV* of the camera, which is the region colored in yellow. Next, the shape's *Minimum Bounding Rectangle* (MBR) is calculated—the geometry colored in green—and used to perform an *R*-tree* spatial search among the GIS source objects. The objects labeled with letters are returned because they overlap with the *MBR*. Among

(a) (b)

Fig. 2. (a) An illustration of a camera's field of view in Google Earth (Copyright © 2013 Google [Google 2013]). (b) The corresponding scene projection on the 2D plane.

the returned objects, A, D, and E appear in between the camera and the landmark, so their height needs to be checked to see if an occlusion occurs. Finally, *GeoLVD* determines the visibility of the landmark after removing all the occlusion situations.

ALGORITHM 1: *GeoLVD* — Geo Landmark Visibility Determination Query Processor

Input:
The geometry of the viewable scene G_{frame}: $FOVScene$;
The geometry of the queried landmark $G_{landmark}$;
The geometry set of all the relevant objects $G_o = \{G_{o_1}, G_{o_2}, \ldots, G_{o_k}\}$;
Output:
The visible angle ranges of the queried landmark $VisibleR$;
$VisibleR = InitVisibleRanges(FOVScene, G_{landmark})$;
$G'_o = GeometryFilter(VisibleR, G_o)$;
for *each* $G_{o_i} \in G'_o$ **do**
  $OccludedRC = ComputeOccludedRC(VisibleR, G_{o_i})$;
  for *each* $r \in OccludedRC$ **do**
    if $!isVerticalOccluded(G_{landmark}, G_{o_i}, r)$ **then**
      $OccludedRC = OccludedRC - \{r\}$;
    end
  end
  $UpdateVR(VisibleR, OccludedRC)$;
  if $VisibleR = \emptyset$ **then**
    break;
  end
end
return $VisibleR$;

To describe the problem formally, let $r = [\mu, \nu]$ denote a horizontal angle range. Then the visible angle ranges of the landmark within the *FOV* can be denoted as a set $VisibleR = \{r_1, r_2, \ldots, r_k\}$. The goal of the algorithm is to compute the visible angle ranges of the queried landmark. A value of $VisibleR = \emptyset$ indicates that the landmark is invisible, otherwise, it is considered visible. Algorithm 1 sketches the overall procedure to determine a landmark's visibility. *InitVisibleRanges()* initializes the horizontal visible angle ranges of the landmark within the FOV without considering occlusions. *GeometryFilter()* filters out the objects that will not cause occlusions by executing an R^**-tree* spatial search with the query area as the MBR of *VisibleR*. All the returned

objects are examined one by one to see if an actual occlusion occurred. *ComputeOccludedRC()* computes the horizontal occluded angle ranges, which are considered as candidates and will be further checked by *isVerticalOccluded()* to see if any of them are vertically occluded as well. If not, the object needs to be removed from the candidate set. *UpdateVR()* updates the visible angle ranges by subtracting the occluded angle ranges at the end of each iteration. The algorithm terminates when $VisibleR$ becomes empty or after all the relevant objects have been checked for potential occlusions.

Next we compare the retrieval performance of the content- and geo-based methods. For this purpose, we first introduce the datasets and the experimental settings that were used.

4. EXPERIMENTAL SETTINGS AND DATASETS

We selected eight popular landmarks in Singapore as our retrieval targets as follows: the Marina Bay Sands hotel, the Esplanade, the Singapore Flyer, the Art Science Museum, the Gardens by the Bay,[2] the Merlion,[2] the Universal Studios Globe[2] and the Ngee Ann City.[2] For each of these landmarks, we collected images from Flickr by posting queries while setting both text and location restrictions. Next, the image sets were manually filtered, keeping only 250 images for each landmark with considerations for both high quality and good diversity. We found that some landmarks might cooccur in the same image (e.g., the Marina Bay Sands hotel and the Art Science Museum). Thus, to reduce complexity, we prepared a common negative examples set for all the landmarks. The images in the negative set were collected from Flickr consisting of other landmark images around the world and images taken in Singapore, and again applying a manual filter and retaining 750 images in the end. As a result, a 1,000-image training set was formed for each of the landmarks including 250 positive and 750 negative instances.

The video collection on which we performed the landmark search consists of 131 geo-referenced videos taken in Singapore. These videos were recorded with smartphone apps that we have developed for both Android and iOS. In these apps the location and orientation metadata is recorded along with each video stream by sampling the GPS sensor every second and the compass sensor every 200 milliseconds. To understand the impact of illumination on any content-based retrieval we further divided the videos into two groups of 114 daytime and 17 nighttime videos. The groundtruth of a landmark's visibility was annotated manually frame by frame at a sample rate of five per second. The groundtruth annotations distinguish the following three situations: (1) landmark entirely visible, (2) partially visible, and (3) invisible. Several examples of frames with fully visible and only partially visible landmarks in the test set are displayed in Figure 3.

The local feature we used were *SIFT* descriptors of 16×16 pixel patches computed over a grid with a spacing of 8 pixels. We adopted a three-level pyramid matching with a vocabulary size of 1,024 for both *ScSPM* and *LLC*.

5. FRAME RETRIEVAL EVALUATION

Each video was treated as a collection of frames and we evaluated the different methods on the task of frame retrieval. The test set for each run was formed by randomly choosing 1,000 frames for querying from the video collection. The proportion of positive to negative samples in the test set was set to 3:7, considering the fact that a landmark usually appears only in a small portion of a video. The retrieval techniques were

[2]These final four landmarks were mainly used in testing the scalability of the proposed hybrid method due to some dataset restrictions: we had limited nighttime videos in the test set and there was a lack of Google StreetView images.

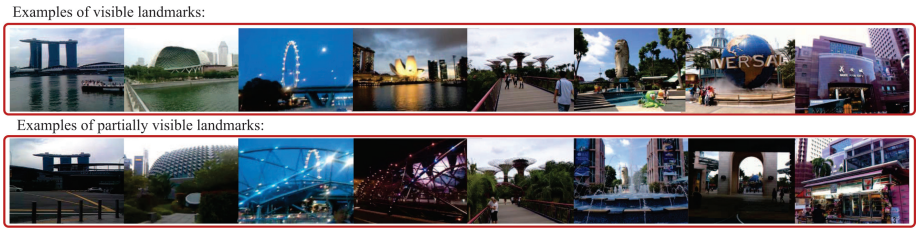


Fig. 3. Illustration of frames with fully and partially visible landmarks in the test set.

Table I. Retrieval Technique Comparison over Different Landmarks and Video Conditions

(a) Marina Bay Sands hotel

Illumination	Day		Night	
Criterion	Precision	Recall	Precision	Recall
ScSPM	**0.8972**	0.5410	**0.7393**	0.4547
LLC	0.8838	0.4927	0.6868	0.4553
GeoLVD	0.7144	**0.8910**	0.6545	**0.6640**

(b) Esplanade

Illumination	Day		Night	
Criterion	Precision	Recall	Precision	Recall
ScSPM	**0.9396**	0.3273	0.6151	0.1347
LLC	0.9099	0.3173	0.6688	0.2113
GeoLVD	0.6991	**0.8450**	**0.6706**	**0.9223**

(c) Singapore Flyer

Illumination	Day		Night	
Criterion	Precision	Recall	Precision	Recall
ScSPM	**0.8845**	0.1100	0.4181	0.0247
LLC	0.7308	0.0587	0.3150	0.0557
GeoLVD	0.7910	**0.8307**	**0.6143**	**0.7343**

(d) Art Science Museum

Illumination	Day		Night	
Criterion	Precision	Recall	Precision	Recall
ScSPM	**0.8936**	0.3223	0.5735	0.3500
LLC	0.8614	0.3113	0.5288	0.4927
GeoLVD	0.7314	**0.8213**	**0.7417**	**0.7450**

Table II. Retrieval Technique Comparison over Supplementary Landmarks among daytime Videos.

Landmark	Gardens by the Bay		The Merlion		Universal Studios		Ngee Ann City	
Criterion	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
ScSPM	**0.8165**	0.5037	**0.9874**	0.2810	**0.8550**	0.4577	**0.9112**	0.1930
LLC	0.7714	0.8597	0.9598	0.5630	0.7760	0.4330	0.6641	0.1630
GeoLVD	0.7070	**0.9740**	0.8183	**0.8967**	0.8367	**0.5870**	0.8612	**0.8537**

evaluated on the criteria of precision, recall, and execution time. Ten experimental runs were carried out and the average results are reported in Tables I, II, and III. Our tests were performed on a desktop computer with a 3.20 GHz dual core CPU and 4GB of main memory. Both the entirely and partially visible landmarks were regarded as positive instances. The classification threshold of the *SVMs* is set to the mid-value of the output range, which is 0.5 when the output is a probability score varying between zero and one. For the spatial index structure we used the Java $R^\star$*-tree* implementation available from https://code.google.com/p/spatialindex.

5.1. Result Discussion

We first report our main observations from Tables I, II, and III and then discuss interesting phenomena and their potential reasons. Finally we summarize the strengths and weaknesses of the content-based and geo-based methods, respectively.

—*Observation 1: Dependence on Illumination.* Among daytime videos, the geobased method always achieves a better recall while the content-based methods achieve better precisions. However, among the nighttime videos the geobased method outperforms the content-based methods in both recall and precision, except for the one case of the Marina Bay Sands hotel.

Table III. Execution Time per Query Frame of the Content- and Geo-Based Methods

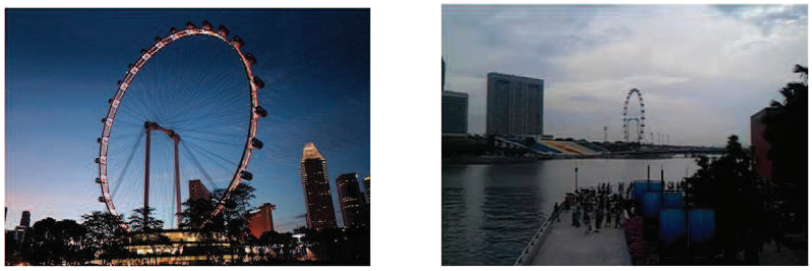
Step	Preprocessing	Retrieval
ScSPM	2.4 s	0.17 ms
LLC	1.1 s	0.17 ms
GeoLVD	-	0.30 ms

—*Observation 2: Dependence on Landmark.* The performance of both *ScSPM* and *LLC* varies significantly among the different landmarks we tested, while the *GeoLVD* method performs relatively stably.
—*Observation 3: Execution Time.* The three methods exhibit comparable, high retrieval speeds, but the content-based methods involve an extra preprocessing step for visual feature extraction and coding, which is time-consuming. Though *LLC* is much faster than *ScSPM* in the preprocessing step, it performs less stable and mostly worse in terms of precision and recall.

It is obvious that the illumination has a great impact on the content-based methods. This is not unexpected because objects become less distinguishable in low light even for human eyes, and therefore also for *SIFT* feature descriptors. On the other hand, illumination should have little impact on the geo-based method, so ideally *GeoLVD*'s performance should be similar on videos taken under various conditions. However, this is not always the case as is illustrated in Table I. For example, the precision of *GeoLVD* for the Singapore Flyer has a gap of 18% between day and night. We consider this to be caused by the variations in the video sets' geographic distributions which also varies, besides the illumination. The location where a video is recorded is a crucial factor for a geobased method because smaller obstacles that may not be stored in a GIS database (e.g., OpenStreetMap), such as trees and unimportant buildings, are more likely to exist in some places than others. Therefore, the retrieval precision could be significantly reduced under such circumstances.

The second observation concerns the impact of the diversity of landmarks. For the content-based methods, though the *SIFT* feature descriptor we used is invariant to uniform scaling, orientation, and affine distortions to some extent, the real-world landmark variations are far more complex than that. In general, frames in which the landmark occupies the majority of the scene can provide more distinguishable local features, so they are easier to be recognized by the classifier. Furthermore, the similarities and differences between the training and test images also affect the classifier's decision. The training set is expected to encompass the landmark's visual diversity in order to well represent all possible situations that may appear in the test images, but as is shown in our experiments, the Flickr images cannot always represent the video frames well. For example, Figure 4 shows two images of the Singapore Flyer. Figure 4(a) is a representative of the Flickr images and Figure 4(b) is a representative of the video frames. Since they were taken at different places, the appearance of the Singapore Flyer differs markedly. We may speculate that there are fewer locations where users take photos of landmarks compared to places where videos are taken, because videos may include actions such as pans and zooms and a landmark may only be part of a lengthy shot. Therefore, in videos recorded at places other than the favorite spots for taking Flickr photos, the landmark becomes more difficult to be recognized by the classifier.

When analyzing the execution time, one key difference between content-based and geobased methods is that the former requires an extra preprocessing step before retrieval can be performed. Visual features need to be extracted and coded in the preprocessing phase for selected frames. Although in video retrieval applications the frame contents are usually analyzed only once and the features saved for indexing and



(a) (b)

Fig. 4. Two typical images of the Singapore Flyer landmark in the experimental dataset: (a) a representative of the Flickr training images (Copyright © 2013 Flickr [Flickr 2013]) and (b) a representative of the frames from the video dataset.

searching, it can usually not be ignored because it is very time-consuming compared with the actual retrieval. In small to medium sized video sharing systems, the time spent on video preprocessing may be acceptable, but as it becomes significant in large video sharing systems such as YouTube, issues arise such as determining the best time to preprocess a video. Though generally only a small portion of videos gain great popularity while the others receive little attention, it still remains a very difficult task to predict the popularity of a video beforehand. It may be wasteful to preprocess a video immediately after it is uploaded because it could turn out to be an unpopular clip that users rarely search for. On the other hand, if preprocessing is performed on demand of a retrieval request, it will cause major delays and possibly make the system impractical. The geo-based method, on the other hand, does not encounter this kind of concern.

5.2. Content- versus Geo-based: Strengths and Weaknesses

As discussed earlier, the content-based method is susceptible to variations in illumination and landmark appearance. It is challenging to form a training set that can well represent all the possible conditions of video frames. Additionally content-based methods require a time-consuming preprocessing step before retrieval. The strength of content-based methods is that an *SVM* classifier outputs a probability. Thus by choosing different thresholds, one can find the best trade-off between precision and recall. The strength of the geo-based method is that it is more stable under various video conditions and landmark appearances. Since videos do not need to be preprocessed, they can be retrieved immediately after being uploaded to a server. Note that although the performance of the geo-based method depends on the accuracy of the sensor data, we found from our experiments that the performance is acceptable, since *GeoLVD* achieves an average recall of greater than 80% (see Table I). For comparison purposes we also computed *ScSPM*'s average recall when it achieves an equal precision to *GeoLVD* and it is only less than 40% among daytime videos. However, the geo-based method has its own drawback. It relies on geographic information services such as OpenStreetMap which may have uneven building and detail coverage, leading to missed occlusions and obstacles. However, if past experience is any guide then these data sources are rapidly improving in both coverage and details. Though this still leaves the problem of dynamic occlusions such as a bus passing by or newly planted trees.

6. CONTENT-BASED METHOD ROBUSTNESS ANALYSIS

From our experiments we have an approximate idea of the performance that can be achieved by *ScSPM* and *LLC* in frame retrievals for certain landmarks. Though these methods have been reported to achieve promising classification results on some standard datasets (e.g., Caltech 101, Caltech 256, and PASCAL VOC 2007) they face

Fig. 5. Details of the Google StreetView training images collected near Marina Bay. Right: image location distribution, left: an example of four side views per location. Copyright © 2013 Google [Google 2013].

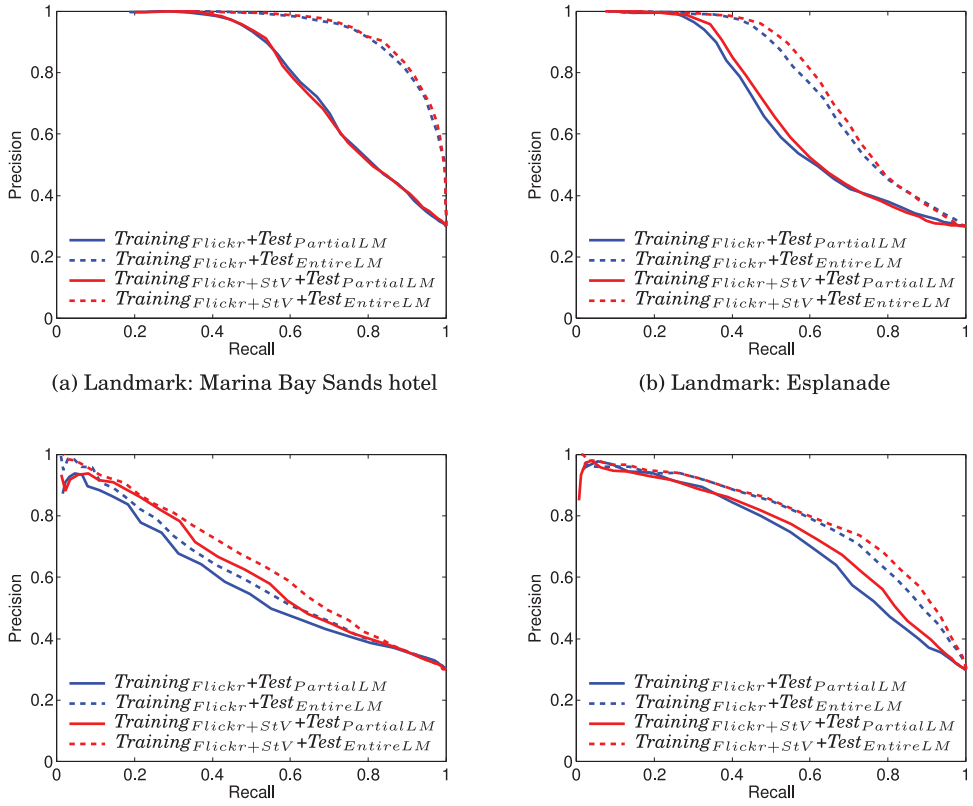
challenges when applied to large and multi-domain real world images and video frames. Here we mainly consider and discuss the following three factors that affect the retrieval effectiveness of content-based methods:

(1) the complexity of visual features,
(2) the representativeness of the training set,
(3) the diversity of the test set.

In general, better performance is more likely achieved by adopting a more extensive spatial model and a larger visual vocabulary, but this will also cost more memory and computation time. The visual feature vector we used for each image has a length of $(1 + 4 + 16) \times 1{,}024 = 21{,}504$ floating point values which is already highly complex for video collections. Since the videos are georeferenced, we expect that their performance can be further improved with the help of the geographic metadata. The integration of content and context information will be discussed in Section 8.

Next, we investigated alternative image sources that can make the training set more representative. Google Maps StreetView would seem a good choice, since it is a very comprehensive dataset which consists of 360° panoramic views of almost all main streets and roads in a number of countries, with a distance of about 12 m between recording locations. To supplement the current training image set, we collected Google StreetView images on five main streets around Marina Bay. In Figure 5, the red pins represent the locations of images we collected, and at each location four directional side views were retrieved, ensuring that one of them would point to the landmark of interest, which is the Singapore Flyer in the case of Figure 5. For each landmark, 25 images were manually selected as positive instances, and the other 75 images with the same location but different side views were automatically selected as negative instances. To evaluate the effectiveness of this new image source, a new training set was formed by randomly selecting 25 positive and 75 negative instances from the previous training set and substituting them with the Google StreetView images. In the following experiment, we will use $Training_{Flickr}$ and $Training_{Flickr+StV}$ to denote the previous and new training sets, respectively.

For the third factor in our list, we conjecture that the appearance of objects may be more diverse in videos than in images because videos frequently contain moving scenes which are different than the composition of single, static picture. Thus, the partial appearance of a landmark is a common situation in video frames. It occurs due to scene transitions, pans, zooms, or partial occlusions. To measure the fraction of a landmark's partial appearances in videos, we filtered out all the frames in which the landmark is only partially visible and prepared a new test set considering only the landmark's



(a) Landmark: Marina Bay Sands hotel

(b) Landmark: Esplanade

(c) Landmark: Singapore Flyer

(d) Landmark: Art Science Museum

Fig. 6. Evaluations with two training sets (Flickr only or Flickr with Google StreetView images, *Flickr+StV*) and two test sets (partial or entire view of landmarks) for the content-based *ScSPM* method.

entire appearance as positive instances. We will use $Test_{PartialLM}$ and $Test_{EntireLM}$ to denote the previous and new test sets respectively in the following experiments.

Experiments were carried out based on different combinations of training and test sets. For the content-based method we used *ScSPM*, and since the Google StreetView images we collected are all taken during the day, we performed the experiments among daytime videos only. As the frames were ranked based on the *SVM* probabilistic outputs, we computed the precision-recall curves by changing the threshold value. The results are shown in Figure 6. We observe that the StreetView images indeed enhance the representativeness of the training set and produce the greatest improvement for the Singapore Flyer landmark. Moreover, when the definition of positive instances is narrowed in the test set to the landmark's entire appearance only, *ScSPM* performs much better than before. This reveals not only *ScSPM*'s poor recognition rate with a landmark's partial appearance, but also the widespread existence of such frames in videos. A landmark is more likely to show up partially if it is bigger in size, which is the reason why the performance on the Marina Bay Sands hotel improves the most.

7. GEO-BASED METHOD ROBUSTNESS ANALYSIS

The geo-based method geometrically computes a landmark's visibility based on the geographic information of both the camera and the surrounding objects (e.g., buildings).

Table IV. Details of Landmark Visibilities in Google StreetView Images

Condition	Visible	Invisible	Total
Marina Bay Sands	298	117	415
Esplanade	120	295	415
Singapore Flyer	233	182	415
Art Science Museum	90	325	415

We summarise the factors that affect the retrieval effectiveness of geo-based methods into the following three aspects:

(1) the accuracy of the smartphone sensor data,
(2) the level of detail of the static geographic object models (e.g., buildings),
(3) the probability of the accidential appearance of real-world dynamic, moving objects such as people, vehicles, *etc.*

To evaluate the influence of each aspect, we used the Google StreetView images for the test set because they are associated with accurate location and orientation information. For each landmark, the test set is formed by all the images whose camera orientation is towards the landmark. We collected StreetView images at 415 locations, so there are overall 415 images in each test set. The details are shown in Table IV. We evaluated two variations of the geo-based method on StreetView images and video frames, respectively. The first approach is *GeoLVD*, which is termed the geo-advanced method as it checks if a landmark is hidden by obstacles. The other is termed the geo-basic method that does not perform the occlusion check, and hence functions for baseline comparisons.

For StreetView images, the geo-basic method always has a recall of 100% because all the test images point toward the landmark and will be surely retrieved without the occlusion check. Its precisions are 71.8%, 28.9%, 56.1%, and 21.7%, respectively, for the four landmarks. Comparatively, the geo-advanced method achieves higher precisions which are 85.2%, 42.0%, 66.5%, and 46.4%. The results are illustrated in Figure 7 and show that the geo-advanced method gains an average increase of 15.4% in precision over geo-basic. Since the StreetView images are associated with accurate geographic information, any retrieval errors of *GeoLVD* are mainly attributed to the second and third factors listed earlier. Geographic information services such as OpenStreetMap only record data of major objects, hence it is not feasible to obtain information of unimportant buildings or trees which are potential obstacles that may hide a landmark. Moreover, even for the buildings recorded in OpenStreetMap, interestingly the height information is mostly not available. For simplicity, we estimated the height according to a building's name and type. For example, we used a height of 0 m for rivers and 30 m for ordinary buildings. Even this simple height estimation as the 3D occlusion check improves the precision by an average of 15.4%. However, it also worsens the recall slightly by an average of 4.3% resulting in a drop from 100% to 95.7%. We conclude that the geographic information collected from OpenStreetMap is currently not detailed enough for precise landmark visibility calculations. As a result the precisions of *GeoLVD* for the Esplanade and Singapore Flyer are still quite low (less than 50%). In general, smaller landmarks are affected more by the lack of information detail.

For video frames the retrieval precisions are visualized in Figure 8. We executed the experiments on 10 test sets for each landmark, and the resulting average precisions are 65.7%, 69.9%, 74.9%, 66.2% for the geo-basic method and 71.4%, 70.0%, 79.1%, 73.1% for the geo-advanced method, respectively. Compared with the results in Figure 7 we can make two major observations. First, both the geo-basic and the geo-advanced methods perform better on video frames, and second the performance gap between

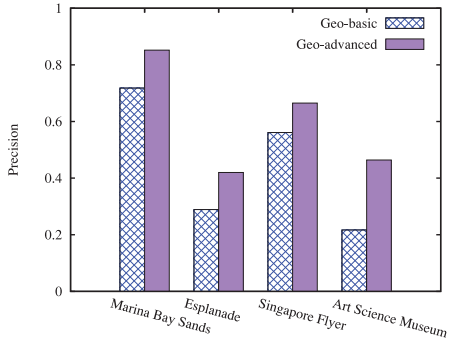


Fig. 7. *GeoLVD* method results with retrieval queries based on Google StreetView images.

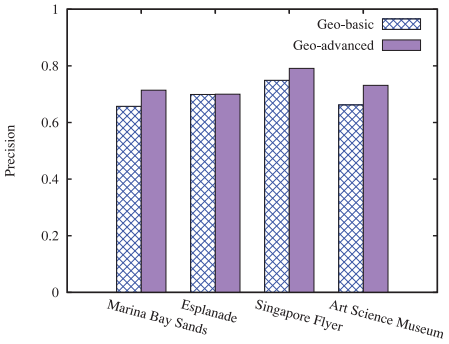


Fig. 8. *GeoLVD* method results with retrieval queries based on video frames.

the geo-basic and the geo-advanced methods is significantly smaller on video frames. Given the assumption that the geographic sensor metadata is not as accurate as the data from StreetView, we would expect the retrieval performance to become worse. One explanation for the improvement might be that the locations for capturing landmarks have been selected by users and hence they have already been "filtered" to avoid occlusions. Hence, compared with the semirobotically collected StreetView images, the probability that a camera view is pointing towards a famous landmark but is occluded by obstacles becomes smaller for video frames. Therefore, even without knowledge of minor inconsequential objects, the geo-based methods can still work well on video frames. The degree of sensor errors can be estimated by the recall values of the geo-basic method, which are 89.6%, 84.5%, 83.1%, and 82.1%, because the values are expected to be 100% if the sensor data is accurate. In terms of the geo-advanced method, only the recall for the Marina Bay Sands hotel decreases slightly to 89.1% while the other three values remain unchanged. This slight drop in recall, together with the other two previous observations, indicate that the geo-based landmark retrieval is less susceptible to the lack of geographic information when queried with video frames.

8. HYBRID INTEGRATED CONTENT AND CONTEXT ANALYSIS

The encouraging results from Section 5 illustrate that contextual information, for example, in the form of geographic data, is a powerful tool for the search of large-scale video archives. Since content- and context-based methods make use of complementary information, it seems natural to combine the two to further improve performance.

Here we propose a hybrid landmark retrieval method which is a late fusion approach that combines the scores of the content- and context-based methods in semantic space [Snoek et al. 2005; Atrey et al. 2010]. As pointed out by Atrey et al. [2010], the late fusion approach has the advantage of allowing us to use the most suitable methods for analyzing each single modality, for instance, *SVMs* for visual content and *GeoLVD* for geographic context. In our study context refers to the location and orientation of the camera. However, the method may be extended to other contextual information in the future. The key idea of the hybrid approach is to first estimate the effectiveness of the content analysis based on the distance from the recorded frames to the landmark, and then use this measure as a weighting factor to combine the scores of the content- and context-based methods. The F1 score is a good indicator for the effectiveness of the content-based method as it considers both precision and recall (see Equation (3)). Therefore, we grouped video frames based on their distance to the landmark and computed the F1 score for every group. We fit the F1 score to a Gaussian function of distance for each of the landmarks as shown in Figure 9. For an image that is taken

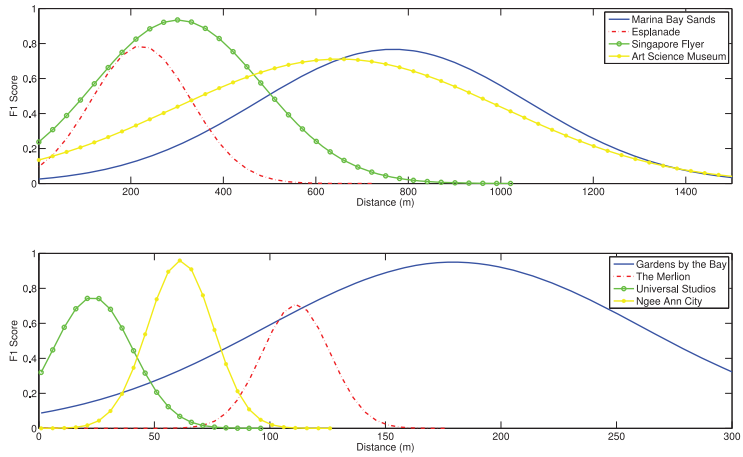


Fig. 9. Estimated Gaussian Function of F1 Score over Distance for different landmarks.

far away from the landmark, the content-based method is not very reliable since the image contains limited details of the landmark but much irrelevant elements from its surroundings. As the camera location moves closer to a landmark, its structure is better visualized and the "noise" from the surroundings is reduced to some extent as well. However, if the distance is further reduced, the probability that an image focuses on only one part of the landmark instead of the whole structure becomes higher, and the resulting information loss also weakens the effectiveness of the content-based method to some extent.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}. \tag{3}$$

We propose to combine the scores of content- and context-based methods, $Score_c$ and $Score_g$, using the following formula:

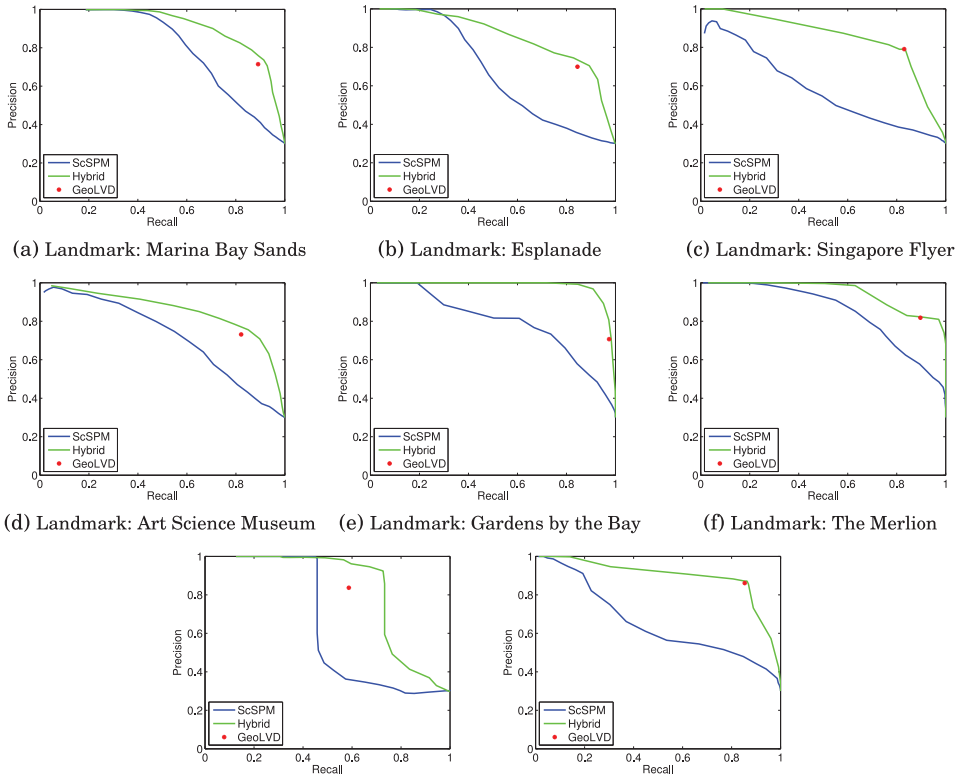
$$Score_h = \alpha \times Score_c + (1 - \alpha) \times Score_g,$$

where α is the preference coefficient that controls the balance between the two scores. Based on the given observations, we define α as

$$\alpha = a_i + b_i \times Gaussian^i_{F1}(Dist(landmark_i, frame)),$$

where $i = 1, 2, \ldots, 8$ represents the index for the eight landmarks, $Gaussian^i_{F1}(x)$ denotes the estimated Gaussian function of the F1 score for landmark i, and $Dist(landmark_i, frame)$ computes the distance between landmark i and a given frame. Values a_i and b_i are constants such that $a_i + b_i \leq 1$. As indicated earlier, we use the F1 score as the measure of effectiveness for the content-based method. In practice, it should be avoided that α is close to zero because the location associated with a frame is acquired from GPS and thus contains some noise. Consequently, we use a_i to control the lower bound of α.

Figure 10 illustrates the precision-recall graphs of the proposed hybrid method as well as the results of *ScSPM* and *GeoLVD* for comparison. The experiments are performed on daytime videos. The positive instances of video frames are defined to include both the entire and partial appearances of a landmark. For each of the landmarks, a separate training set is selected where a_i and b_i $(i = 1, 2, \ldots, 8)$ are tuned. We also illustrate the highest F1 score that each of the methods achieves in Table V.



(a) Landmark: Marina Bay Sands (b) Landmark: Esplanade (c) Landmark: Singapore Flyer

(d) Landmark: Art Science Museum (e) Landmark: Gardens by the Bay (f) Landmark: The Merlion

(g) Landmark: Universal Studios (h) Landmark: Ngee Ann City

Fig. 10. Precision-recall curves comparison of methods based on content only, context only, and their hybrid integration.

Table V. F1 Scores of Methods Based on Content Only, Context Only, and Their Hybrid Integration

Method	*ScSPM*	*GeoLVD*	*Hybrid*
Marina Bay Sands	0.6858	0.7930	0.8249
Esplanade	0.5736	0.7652	0.7878
Singapore Flyer	0.5628	0.8104	0.8104
Art Science Museum	0.6783	0.7737	0.8003
Gardens by the Bay	0.7347	0.8193	0.9376
The Merlion	0.7451	0.8557	0.8836
Universal Studios	0.6261	0.6900	0.8138
Ngee Ann City	0.6169	0.8574	0.8701

From the statistics we can see that the proposed hybrid method achieves the best result overall. It significantly improves the effectiveness of *ScSPM*. Moreover, it enhances *GeoLVD* not only by increasing the F1 score but also by enriching the output from binary to a probability range. The best improvement is observed with the landmark Universal Studios Globe. We examined the test set and found that the frames are usually recorded very close to the globe because of its small size compared with other landmarks, and this therefore makes it the most sensitive to GPS errors. This is also the reason why *GeoLVD* only achieves an average recall of 58.7% with the globe. Fortunately, *ScSPM* compensates for such situations very well and consequently makes the

hybrid method highly effective. On the other hand, most of the test frames from the Singapore Flyer are taken some distance away at Marina Bay which weakens the reliability of *ScSPM* in terms of compensation. The execution time of the hybrid method is approximately the sum of the content- and context-based methods we used, that is, 2.4 s for preprocessing and 0.47 ms for retrieval per frame, which is quite fast and acceptable. In summary, the selection of a suitable method depends on the application requirements of retrieval effectiveness which is closely related to the characteristics of the landmarks and the video collection.

9. CONCLUSIONS AND FUTURE WORK

In this study we have evaluated two state-of-the-art video landmark retrieval paradigms, namely media-content based and geocontext based retrievals. For the content-based retrieval, we selected two high performance methods, *ScSPM* and *LLC*, and for the geo-based retrieval, we introduced *GeoLVD*, which is inspired by the 3D visibility query processor proposed by Shen et al. [2011].

From the comparison results we draw a number of interesting observations, chiefly among them is the importance of the illumination conditions for content-based methods, summarized as follows.

—When performing retrievals from our daytime video collection, it is always the case that content-based retrieval achieves a higher precision, while the geo-based retrieval achieves a higher recall.
—However, when carrying out retrievals from the nighttime video collection, the geo-based method always outperforms the content-based method in terms of both precision and recall.
—The performance of the content-based method varies significantly when searching for different landmarks, while the geo-based method is relatively stable.

Therefore, we conclude that when the illumination or the appearance of a landmark is not favorable for content-based retrieval, the geo-based method is more suitable and should be chosen instead.

In terms of execution time, we observe the following.

—Both the content-based and the geo-based methods exhibit comparable retrieval speeds in the sub-milliseconds per frame, but the former involves an extra preprocessing step for visual feature extraction and coding which is usually time-consuming, taking on the order of 1 to 2 seconds per frame.
—*LLC*, which aims at speeding up the visual feature coding procedure, is much faster than *ScSPM* in the preprocessing step. However, it is also less accurate in the retrieval phase.

The time spent on the preprocessing step may not be a significant burden for small to medium video sharing systems. However for large video sharing platforms such as YouTube, it becomes a hassle to determine when is the best time to preprocess a video. In such a scenario the geo-based retrieval method reveals its strength as no video preprocessing is needed beforehand. However, the standardized recording of geographic metadata is currently still in the exploratory stage among large online media sharing systems.

In the future, we plan to extend the evaluation from frame retrieval to segment retrieval by investigating video temporal continuity as well. Additionally, efforts have been made in the study and development of indoor positioning systems, for instance, the Redpin project (http://redpin.org). We are interested in integrating such capabilities into our system to allow it to work both outdoors and indoors.

REFERENCES

Giuseppe Amato, Fabrizio Falchi, and Fausto Rabitti. 2012. Landmark recognition in VISITO Tuscany. In *Multimedia for Cultural Heritage*, 1–13.

Sakire Arslan Ay, Roger Zimmermann, and SeonHo Kim. 2010. Relevance ranking in georeferenced video search. *Multimedia Syst.* 16, 2, 105–125.

Sakire Arslan Ay, Roger Zimmermann, and Seon Ho Kim. 2008. Viewable scene modeling for geospatial video search. In *Proceedings of the ACM International Conference on Multimedia*. 309–318.

Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: A survey. *Multimedia Syst.* 16, 6, 345–379.

Yannis Avrithis, Yannis Kalantidis, Giorgos Tolias, and Evaggelos Spyrou. 2010. Retrieving landmark and non-landmark images from community photo collections. In *Proceedings of the ACM International Conference on Multimedia*. 153–162.

D. M. Chen, G. Baatz, K. Koser, S. S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, Xin Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. 2011a. City-scale landmark identification on mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 737–744.

Tao Chen, Kim-Hui Yap, and L.-P. Chau. 2011b. Integrated content and context analysis for mobile landmark recognition. *IEEE Trans. Circuits Syst. Video Technol.* 1476–1486.

O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. 2007. Total recall: automatic query expansion with a generative feature model for object retrieval. In *Proceedings of the International Conference on Computer Vision*. 1–8.

G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. 2004. Visual categorization with bags of keypoints. In *Proceedings of the ECCV International Workshop on Statistical Learning in Computer Vision*. 1–22.

Shaolei Feng and R. Manmatha. 2008. A discrete direct retrieval model for image and video retrieval. In *Proceedings of the International Conference on Content-Based Image and Video Retrieval*. 427–436.

Efstratios Gavves, Cees G. M. Snoek, and Arnold W. M. Smeulders. 2012. Visual synonyms for landmark image retrieval. *Comput. Vision Image Understand* 116, 12, 238–249.

Qiang Hao, Rui Cai, Zhiwei Li, Lei Zhang, Yanwei Pang, and FengWu. 2012. 3D visual phrases for landmark recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3594–3601.

E. Hecht. 2001. *Optics* (4th ed.). Addison-Wesley Publishing Company.

N.V. Hoàng, V. Gouet-Brunet, M. Rukoz, and M. Manouvrier. 2010. Embedding spatial information into image content description for scene retrieval. *Pattern Recog.* 43, 9, 3013–3024.

Zi Huang, Bo Hu, Hong Cheng, Heng Tao Shen, Hongyan Liu, and Xiaofang Zhou. 2010. Mining near-duplicate graph for cluster-based reranking of web video search results. *ACM Trans. Inf. Syst.* 22:1–22:27.

Ramesh Jain and Pinaki Sinha. 2010. Content without context is meaningless. In *Proceedings of the ACM International Conference on Multimedia*. 1259–1268.

Pascal Kelm, Sebastian Schmiedeke, and Thomas Sikora. 2011. A hierarchical, multi-modal approach for placing videos on the map using millions of Flickr photographs. In *Proceedings of the ACM Workshop on Social and Behavioural Networked Media Access*. 15–20.

Lyndon S. Kennedy and Mor Naaman. 2008. Generating diverse and representative image search results for landmarks. In *Proceedings of the International Conference on World Wide Web*. 297–306.

Youngwoo Kim, Jinha Kim, and Hwanjo Yu. 2012. GeoSearch: Georeferenced video retrieval system. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*. 1540–1543.

Yin-Hsi Kuo, Wen-Huang Cheng, Hsuan-Tien Lin, and Winston H. Hsu. 2012. Unsupervised semantic feature discovery for image object retrieval and tag refinement. *IEEE Trans. Multimedia,* 1079–1090.

S. Lazebnik, C. Schmid, and J. Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2169–2178.

Zhen Li and Kim-Hui Yap. 2012. Content and context boosting for mobile landmark recognition. *Signal Process. Lett.* 459–462.

Xiaotao Liu, Mark Corner, and Prashant Shenoy. 2005. SEVA: sensor-enhanced video annotation. In *Proceedings of the ACM International Conference on Multimedia*. 618–627.

David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision,* 91–110.

Jiebo Luo, Dhiraj Joshi, Jie Yu, and Andrew Gallagher. 2011. Geotagging in multimedia and computer vision—a survey. *Multimedia Tools Appl.* 51, 1, 187–211.

Otávio A. B. Penatti, Fernanda B. Silva, Eduardo Valle, Valerie Gouet-Brunet, and Ricardo da S. Torres. 2014. Visual word spatial arrangement for image retrieval and classification. *Pattern Recognit.* 705–720.

Otávio A. B. Penatti, Lin Tzy Li, Jurandy Almeida, and Ricardo da S. Torres. 2012. A visual approach for video geocoding using bag-of-scenes. In *Proceedings of the ACM International Conference on Multimedia Retrieval*. 1–8.

Adam Rae, Vannesa Murdock, Pavel Serdyukov, and Pascal Kelm. 2011. Working Notes for the Placing Task at MediaEval 2011.

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 International Conference on Multimedia*. 93–102.

Rainer Simon and Peter Fröhlich. 2007. A mobile application framework for the geospatial web. In *Proceedings of the International Conference on World Wide Web*. 381–390.

Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. 2005. Early Versus Late Fusion in Semantic Video Analysis. In *Proceedings of the ACM International Conference on Multimedia*. 399–402.

Fabrice Souvannavong, Bernard Merialdo, and Benoit Huet. 2005. Region-based video content indexing and retrieval. In *Proceedings of the International Workshop on Content-Based Multimedia Indexing*. 21–23.

Xinmei Tian, Linjun Yang, Jingdong Wang, Yichen Yang, Xiuqing Wu, and Xian-Sheng Hua. 2008. Bayesian video search reranking. In *Proceedings of the ACM International Conference on Multimedia*. 131–140.

Ville Viitaniemi and Jorma Laaksonen. 2008. Experiments on selection of codebooks for local image feature histograms. In *Visual Information Systems, Web-Based Visual Information Search and Management*, 126–137.

Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, T. Huang, and Yihong Gong. 2010. Locality-constrained linear coding for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3360–3367.

Jianchao Yang, Kai Yu, Yihong Gong, and T. Huang. 2009. Linear spatial pyramid matching using sparse coding for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1794–1801.

Kim-Hui Yap, Tao Chen, Zhen Li, and Kui Wu. 2010. A comparative study of mobile-based landmark recognition techniques. *IEEE Intell. Syst.* 25, 1, 48–57.

Bo Zhang, Qinlin Li, Hongyang Chao, Bill Chen, Eyal Ofek, and Ying-Qing Xu. 2010. Annotating and navigating tourist videos. In *Proceedings of the International Conference on Advances in Geographic Information Systems*. 260–269.

Yan-Tao Zheng, Ming Zhao, Yang Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven. 2009. Tour the world: Building a web-scale landmark recognition engine. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1085–1092.

Received June 2013; revised November 2013, April 2014; accepted September 2014