3D Visual Phrases for Landmark Recognition

Qiang Hao[1]*, Rui Cai[2], Zhiwei Li[2], Lei Zhang[2], Yanwei Pang[1], Feng Wu[2]
[1]Tianjin University, Tianjin 300072, P.R. China
[2]Microsoft Research Asia, Beijing 100080, P.R. China
[1]{qhao,pyw}@tju.edu.cn, [2]{ruicai,zli,leizhang,fengwu}@microsoft.com

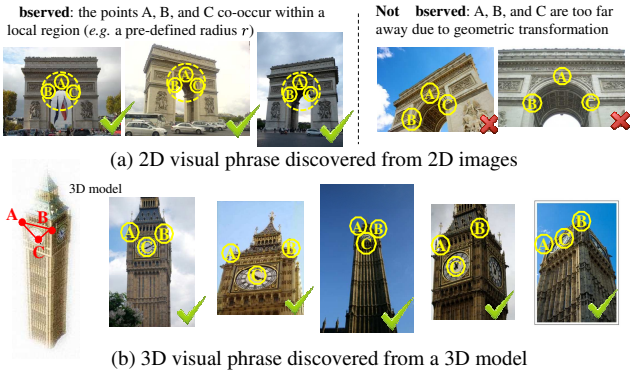
Abstract

In this paper, we study the problem of landmark recognition and propose to leverage 3D visual phrases to improve the performance. A 3D visual phrase is a triangular facet on the surface of a reconstructed 3D landmark model. In contrast to existing 2D visual phrases which are mainly based on co-occurrence statistics in 2D image planes, such 3D visual phrases explicitly characterize the spatial structure of a 3D object (landmark), and are highly robust to projective transformations due to viewpoint changes. We present an effective solution to discover, describe, and detect 3D visual phrases. The experiments on 10 landmarks have achieved promising results, which demonstrate that our approach provides a good balance between precision and recall of landmark recognition while reducing the dependence on post-verification to reject false positives.

1. Introduction

Landmark recognition has become an active research topic in the last few years. The problem is, given a query photo, how to automatically determine where it was taken based on a massive image database collected from the Web. Landmark recognition has great potentials in applications like geo-localization [8, 11, 17] and tourist guide [19, 25].

Most existing work considers landmark recognition as an image retrieval task, and is mainly based on the well-known bag-of-words (BoW) framework [18, 13]. That is, both the query and database images are characterized by local features like SIFT [12], and the database images are indexed to quickly identify similar images matched with the query. Although such a framework is quite efficient in search, it still has limitations. First, the database index contains considerable features extracted from irrelevant objects (*e.g.,* faces, trees, and road markings) [9]. These noisy features inevitably confuse the matching process and hurt the recognition performance. Second, the index construction usually treats the database images independently and thus ignores the geometric relationships between them.



(a) 2D visual phrase discovered from 2D images

(b) 3D visual phrase discovered from a 3D model

Figure 1. 2D visual phrase *vs.* 3D visual phrase. (a) 2D visual phrases are basically co-occurrence statistics within some local regions, which are sensitive to viewpoint changes. By contrast, (b) 3D visual phrases are more reliable since they are derived from the physical structure of a landmark.

Benefiting from the recent progress in large-scale structure-from-motion (SfM) [19, 20, 21], it is no longer an obstacle to reconstruct 3D models from unordered images. Such a 3D model can not only distinguish the target object (landmark) from noisy background, but also keep the geometric relationships between database images. Hence, leveraging 3D models for object recognition has been a noticeable trend. Some approaches adopt 3D models to select or synthesize images taken from iconic viewpoints for 2D image matching [8, 14], while some other approaches directly perform matching between 3D object models and 2D query images [15, 1, 11, 16].

The matching process, either image-to-image or 3D-model-to-image, is essentially finding correspondences between 2D local features. However, the limited discriminative power of individual features usually leads to inaccurate correspondences. Therefore, geometric verification, *e.g.,* planar homography estimation (for image-to-image) or pose estimation (for 3D-model-to-image), is a necessary post-processing step so as to eliminate false matches. To re-

*This work was performed at Microsoft Research Asia.

duce the dependence on geometric verification and relieve the ambiguity of individual features, people either try to embed local context into visual words [4, 5], or construct visual phrases [22, 23, 24] or feature triplets [26] to enhance distinctiveness. All these attempts work on image-to-image matching, and therefore mainly consider contextual information in 2D image planes. For example, as shown in Fig. 1 (a), a (2D) visual phrase is basically a combination of several visual words which frequently co-occur with each other within some local neighborhood areas.

Although visual phrases and extensions have achieved success in many applications, it is non-trivial to define an appropriate local neighborhood area in 2D image planes. The most common practice is to pre-define a radius based on heuristic rules, as shown in Fig. 1 (a). Unfortunately, such a 2D local region inherently lacks robustness to viewpoint changes which serve as a common challenge of landmark recognition. Therefore, 2D visual phrases are not guaranteed to be perspective invariant. This observation motivates us to discover more reliable visual phrases based on 3D models. Instead of finding groups of local features (visual words) in 2D images, we work on identifying groups of 3D points, named as *3D visual phrases*, in the real world space. Such a 3D visual phrase characterizes the intrinsic physical structure of an object (landmark), and is robust to viewpoint changes. As shown in Fig. 1 (b), the three 3D points on Big Ben would co-occur in images taken from various viewpoints where they are simultaneously visible.

Inspired by existing 3D model-based methods, in this paper, we propose to discover 3D visual phrases from 3D landmark models and treat landmark recognition task as the identification of these visual phrases from unseen images. To this end, there are several problems to be answered:

- First, how to construct a reasonable set of 3D visual phrases for a landmark? As the 3D point cloud of a landmark model could contain tens of thousands of points, arbitrary combination is obviously infeasible. Moreover, the discovered visual phrases should provide a comprehensive description to the whole landmark structure.
- Second, how to describe a 3D visual phrase and detect it in unseen images? Both the visual appearance of 3D points and their geometric relationships should be properly characterized. Besides, an efficient scheme is necessary for fast phrase detection from unseen images.
- Third, how to balance precision and recall of recognition? While the preserved geometric constraints tend to provide high precision, 3D visual phrases need properly relaxed detection method to ensure recall on unseen data.

To make the idea of 3D visual phrase-based landmark recognition practical, in this paper, we propose a series of solutions to address the aforementioned problems. First, in Section 2, we define 3D visual phrases based on the triangular facets which approximately cover the surface of a 3D landmark model. Each facet captures a local spatial structure of the 3D model, and all the facets as a whole provide a full description to the landmark. Moreover, the number of facets is with the same order of magnitude as the 3D points. Second, in Section 3 we introduce a comprehensive set of descriptions, for both visual appearance and geometric structure, to characterize a 3D visual phrase. In this part, we try to preserve as much information as possible to ensure the recall of visual phrases on unseen data while keeping sufficient constraints to filter out false detections. Then in Section 4, we propose an efficient detection algorithm to identify 3D visual phrases from a query image. The principal idea is to leverage the spatial correlations between 3D visual phrases as a constraint to quickly reject false positives. Finally, we report the evaluation results in Section 5, and conclude the paper in Section 6.

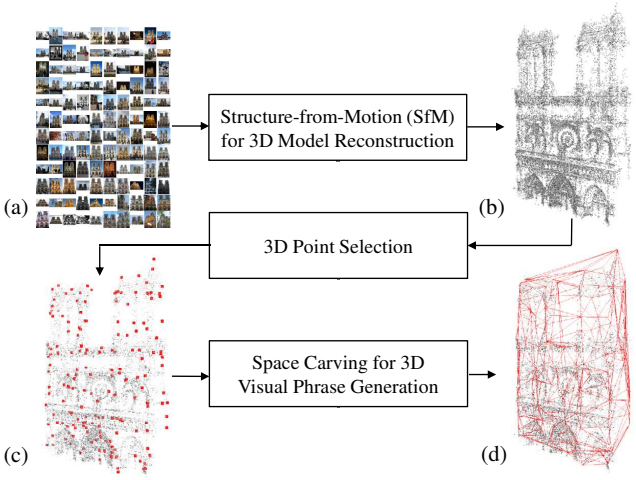


Figure 2. An illustration of 3D visual phrase discovery. Given a set of landmark images in (a), a 3D point cloud (b) is reconstructed using structure-from-motion. From the point cloud, a subset of points (c) are carefully selected, based on which space carving is performed to generate surface facets (d) as 3D visual phrases.

2. 3D Visual Phrase Discovery

In this section, we introduce the discovery of 3D visual phrases from landmark images, as shown in Fig. 2. First, a 3D point cloud is reconstructed from an image collection. Then, the 3D points are sub-sampled while keeping a spatial coverage of the entire model. Finally, a 3D surface is constructed based on the selected 3D points and triangular facets on the surface are harvested as 3D visual phrases.

3D Landmark Reconstruction. Given a set $\mathcal{I}$ of images for a landmark, we use structure-from-motion (SfM) [7] to reconstruct the 3D model [1] . A 3D model here is actually a point cloud composed of a set $\mathcal{P}$ of 3D points, as shown in Fig. 2 (b). Meanwhile, we also obtain the estimated cam-

[1] In practice, it is convenient to perform 3D reconstruction using off-the-shelf tools such as Bundler [19, 20] and VisualSFM [21].

era pose of each image, under the same world coordinate system where the point cloud is defined. For each 3D point $p \in \mathcal{P}$, denote $\mathcal{I}_p \subset \mathcal{I}$ the set of images in which p is observed and registered to the 3D model. Accordingly, the popularity of p is defined as the cardinality of $\mathcal{I}_p$.

3D Point Selection. A point cloud reconstructed from approximately one thousand images typically contains tens of thousands of 3D points, which are usually redundant for landmark recognition. Furthermore, the points may vary a lot in popularity, ranging from two to hundreds of images, and thus have different repeatability in unseen images. Therefore, we need to select a subset of points most important for identifying landmarks, by considering two criteria for point selection, namely (1) *point popularity* and (2) *spatial coverage* of the landmark model. Note that we explicitly require the selected points to cover the 3D model rather than the empirically observed 2D images [11], in order to better handle images taken from arbitrary viewpoints.

To sample a point cloud $\mathcal{P}$ at a given sampling rate η, we first construct an octree to represent the 3D bounding cube of $\mathcal{P}$, and then iteratively partition the most dense (*i.e.*, containing the most points) voxel until obtaining $\eta \times |\mathcal{P}|$ non-empty voxels. Finally, the most popular point in each of these voxels is selected to compose a point subset $\mathcal{P}_\eta$, as illustrated in Fig. 2 (c).

3D Visual Phrase Generation. Taking each point in $\mathcal{P}_\eta$ as a 3D visual word, the most intuitive method for generating a 3D visual phrase is to combine several points into a group. However, arbitrary combination has two major drawbacks. For one thing, it inevitably suffers from the large amount of points, which result in numerous possible combinations intractable to select. For another, it lacks principle for generating a compact and sufficient set of visual phrases for landmark recognition from new images, due to the difficulty in fully covering the landmark structure. Therefore, we propose to derive visual phrases from the facets that compose the surface of a 3D landmark model. The advantages are three-fold: (1) each facet corresponds to a local spatial structure of the landmark and consequently has sound visibility and repeatability in unseen images, (2) the facets as a whole approximate the 3D surface and provide a full coverage of the landmark, and (3) the facets construct a compact set with the same order of magnitude as the 3D points.

For 3D modeling, it is common to approximate an object surface using a number of Delaunay triangles. Motivated by this, we first conduct Delaunay Triangulation on the convex hull that envelopes a given 3D point set, and then refine the surface using space carving techniques [3, 10], *i.e.*, iteratively removing false facets which occlude any points visible from empirical images. Finally, each facet on the resulting surface is taken as a 3D visual phrase, as shown in Fig. 2 (d), which is essentially a triplet of 3D points with a particular geometric structure.

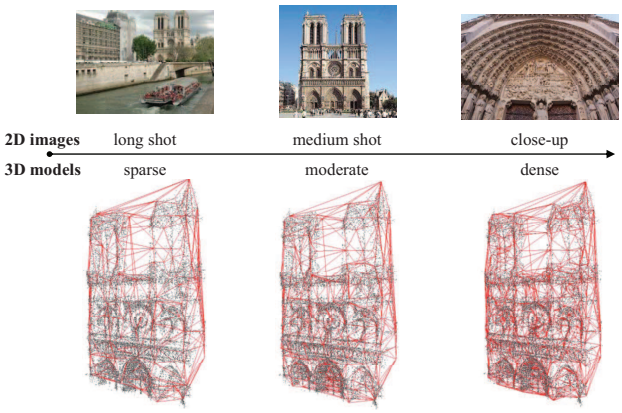


Figure 3. An illustration of multi-scale (sparse, moderate, and dense) 3D visual phrases. Multiple sets of 3D visual phrases are generated from different subsets of 3D points to handle different object sizes (long shot, medium shot, and close-up) in 2D images.

Multi-Scale 3D Visual Phrases. A landmark can be photographed from different distances with different focal lengths, resulting in various sizes of objects in photos. Meanwhile, each facet on the landmark is visible only in a proper scale range. For instance, a small facet may be too local to be identified in a long shot image; while a large facet cannot completely appear in a close-up image where the landmark is only partially observed.

To identify landmarks in different 2D sizes, we need to generate multi-scale 3D visual phrases (facets). Intuitively, the fewer number of points selected, the larger the facets generated. Therefore, we simply select subsets of points from the raw point cloud at various sampling rates (*e.g.*, $\eta = 1\%, 2\%, 4\%$) to generate multi-scale visual phrases, as the *sparse*, *moderate*, and *dense* models shown in Fig. 3, which are created to handle photos taken from *long shot*, *medium shot*, and *close-up*, respectively.

The multi-scale visual phrases can be considered as several models to describe the same landmark at different granularity levels. Without loss of generality, in the subsequent sections, we just introduce how to deal with visual phrases under the same scale.

3. 3D Visual Phrase Description

The discovered 3D visual phrases are expected to be recognizable in unseen images. Such a detection procedure highly relies on recorded characteristics of visual phrases to distinguish true occurrences from false positives. Since each 3D visual phrase is a triangular facet with three vertex points, it is characterized from two perspectives, namely (1) *visual appearance* of each vertex point and (2) *geometric structure* among the points. During visual phrase detection, appearance provides relaxed criteria to recall true positives, while geometric structure serves as constraints to eliminate false positives and boost the precision.

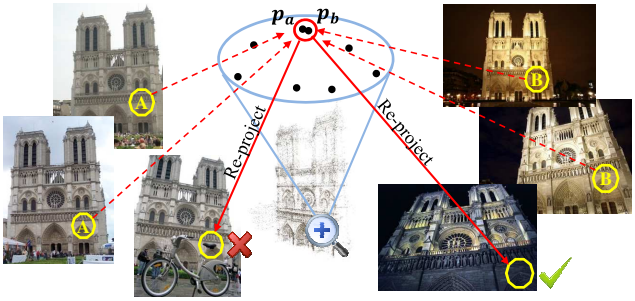


Figure 4. An illustrative example of the motivation for enriching appearance descriptions by 3D-to-2D re-projection. Please refer to the text for more details.

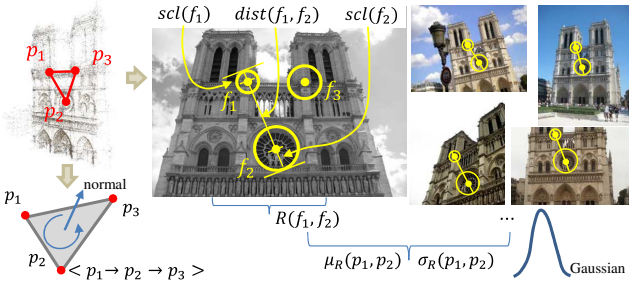


Figure 5. Geometric structure descriptions of a 3D visual phrase: (a) cyclic order and (b) scale-distance cross-ratio.

Visual Appearance. By describing the visual appearance of a 3D point p, we aim to preserve a comprehensive and compact representation, for matching with occurrences of p in unseen images. In this work, we adopt SIFT [12] feature as the appearance description to keypoints in images; each SIFT feature f is characterized by a 128-dimensional descriptor $des(f)$ extracted from a local image patch in scale $scl(f)$. Each 3D point $p \in \mathcal{P}$ appears in a set $\mathcal{I}_p$ of images and accordingly matches with a set $\mathcal{F}_p$ of SIFT features (one per image).

A straightforward appearance description of p is the mean descriptor averaged over SIFT features in $\mathcal{F}_p$ [11]. However, the raw $\mathcal{F}_p$ is usually insufficient to provide a comprehensive description, because 3D reconstruction relies on very strict point matching which inevitably leads to two typical kinds of mismatches. On one hand, one physical object point would be over-split into multiple 3D points close to each other (*e.g.*, p_a and p_b in Fig. 4) during SfM, if the corresponding SIFT features fail to match due to descriptor variation under viewpoint or illumination changes. Obviously, spatially sub-sampling the point cloud as aforementioned would not preserve all such 3D points, and inevitably harm the capability in matching with new images. On the other hand, a 3D point may be visible in some images but the corresponding SIFT features are not registered to the 3D model due to appearance variation.

To compensate for the above loss in the observations of 3D points' appearance, we propose to expand $\mathcal{F}_p$ by re-projecting each 3D point p to all the 2D images in which it is visible, based on the camera poses estimated during SfM. Then, any SIFT feature is appended to $\mathcal{F}_p$ if it is sufficiently close (*e.g.*, with a maximum deviation of two pixels) to such a re-projected 2D position.

To describe the appearance of p compactly, we further compress the expanded $\mathcal{F}_p$ by clustering the descriptors and preserving only a set $\mathcal{A}_p$ of mean descriptors calculated from the $L = 5$ largest clusters. As a byproduct, occasional noise arising from re-projection (*e.g.*, descriptors extracted on occluding objects like the bicycle in Fig. 4) could be removed from the compact description.

Geometric Structure. As aforementioned, comprehensive appearance descriptions to 3D points could enable relaxed point matching and thus suggest abundant candidates when detecting visual phrase from new images. To distinguish true visual phrases from false ones, it is highly desired to preserve inter-point geometric structures as (1) constraints for rejecting false candidates which violate particular geometric properties and (2) criteria for ranking candidates by consistency to stable statistics from the observed data.

Given a 3D visual phrase containing three vertex points, denoted as $\mathbf{v} = (p_1, p_2, p_3)$, a simple and robust geometric property is the cyclic order of the points. Thus, we define a direction around the perimeter of $\mathbf{v}$ based on its outer-pointing surface normal and the right-hand rule; along this direction, a cyclic order of vertex points is defined accordingly. As the example shown in Fig. 5 (a), the point order is $o_{\mathbf{v}} = \langle p_1 \rightarrow p_2 \rightarrow p_3 \rangle$. Such an order is invariant to projective transformations as it holds for any 2D projection (*i.e.*, a triplet of SIFT features). Thus, it provides a rigid criterion for identifying true occurrences of the visual phrase.

Besides the rigid point orders, we consider a more flexible geometric property, by assuming a close correlation between the scales and pairwise distances of 3D points' 2D projections from different viewpoints. Specifically, we define a measure on projections of two 3D points (particularly on two SIFT features with their scales providing two auxiliary points). Let SIFT features f_1 and f_2 be projections of 3D points p_1 and p_2 in an image, respectively; their 2D distance is $d_{1,2} = dist(f_1, f_2)$ and respective scales are $scl(f_1)$ and $scl(f_2)$, based on which a ratio measure is defined as

$$R(f_1, f_2) \triangleq \frac{(d_{1,2} + scl(f_1)) \times (d_{1,2} + scl(f_2))}{d_{1,2} \times (d_{1,2} + scl(f_1) + scl(f_2))}.$$

Such a measure is an analogue of the projective-invariant *cross-ratio* of four collinear points, and thus is robust to viewpoint changes. To tolerate errors in SIFT scale estimation under projective transformations, we average the R values over all images in which both the points are observed, yielding a Gaussian distribution with mean $\mu_R(p_1, p_2)$ and standard deviation $\sigma_R(p_1, p_2)$, as illustrated in Fig. 5 (b).

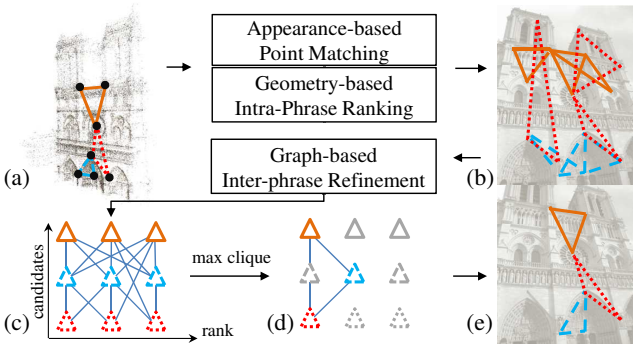


Figure 6. An illustration of the detection of 3D visual phrases from a query image. Please refer to the text for more details.

4. 3D Visual Phrase Detection

Given a set $\mathcal{V}$ of 3D visual phrases and a query image I_q, the landmark recognition task is essentially the detection of 3D visual phrases appearing in I_q, as illustrated in Fig. 6.

Appearance-based Point Matching. In this step, we aim to match every 3D point involved in $\mathcal{V}$ with the SIFT features extracted from I_q based on visual appearance. To ensure the recall of truly appearing 3D visual phrases, we seek for multiple SIFT features as candidates for each 3D point, because the most visually similar ones are not necessarily the real projections of 3D points.

Remind that we have characterized the appearance of each 3D point p as a set $\mathcal{A}_p$ of (mean) SIFT descriptors; thus, the appearance similarity of a SIFT feature f to p can be defined based on cosine similarity as

$$sim_{app}(f;p) = \max_{a \in \mathcal{A}_p} \cos\left(des(f), a\right).$$

For each point, we preserve up to top N (empirically set to 3) candidate SIFT features with appearance similarity above an empirical threshold $\alpha = 0.8$, to control the trade-off between the recall of visual phrases and the scale of solution space. In the implementation, the point matching is accelerated by approximate nearest neighbor search [2].

Geometry-based Intra-Phrase Ranking. After matching 3D points with SIFT features, we can obtain a set $\mathcal{V}'$ of visual phrases that all three vertex points have matched with at least one feature. Since there are usually multiple candidate features per point, each visual phrase in $\mathcal{V}'$ could have dozens of candidates (as in Fig. 6 (b)), in which at most only one is true. To boost the signal-to-noise ratio for subsequent processing, we resort to the known geometric structure of each visual phrase to filter its corresponding candidates and eliminate the false ones with different structures.

Let visual phrase $\mathbf{v} = (p_1, p_2, p_3)$ correspond to a set $\mathcal{C}$ of candidates, each being a triplet of SIFT features, denoted as $\mathbf{c} = (f_1, f_2, f_3)$, where f_i is a candidate projection of p_i. To filter $\mathcal{C}$, we first compare the cyclic order of features in each candidate with the standard order $o_{\mathbf{v}}$, and discard any candidate with a different order. For each remaining candidate $\mathbf{c}$, a geometric similarity score is then calculated to rate the degree to which $\mathbf{c}$ has a similar scale-distance correlation with $\mathbf{v}$, as

$$sim_{geo}(\mathbf{c};\mathbf{v}) = \exp(-\tau \sum_{(i,j)\in\mathcal{E}} \frac{(R(f_i,f_j) - \mu_R(p_i,p_j))^2}{\sigma_R^2(p_i,p_j)}),$$

where τ is a positive coefficient empirically set to 0.1; $\mathcal{E} = \{(1,2),(2,3),(3,1)\}$ enumerates the point pairs. Finally, the candidates are ranked by the overall similarity

$$sim_{overall}(\mathbf{c};\mathbf{v}) = sim_{geo}(\mathbf{c};\mathbf{v}) \times \prod_{i=1}^{3} sim_{app}(f_i;p_i),$$

which serves as the confidence that candidate $\mathbf{c}$ is a true occurrence of $\mathbf{v}$. In the implementation, the candidate ranking list of each visual phrase is truncated to retain at most top $M = 5$ candidates with confidence exceeding $\beta = 0.2$.

Graph-based Inter-Phrase Refinement. A large proportion of false visual phrase candidates can be filtered out by geometric criteria as above. The remaining spurious candidates could be further rejected by considering relationships between candidates for different visual phrases. To this end, we build an undirected graph (as shown in Fig. 6 (c)), in which each node is a visual phrase candidate, and each edge exists if and only if the two candidates it connects could be true simultaneously; such co-occurrence is feasible when two candidates meet both of the following criteria:

- They must NOT lead to ambiguity in point-to-feature matching (*i.e.*, one 3D point matching with multiple features or vice versa).
- As 2D triangles, they must NOT have overlapping coverage areas. This condition always holds for true projections of 3D visual phrases since they are components of a non-overlapping coverage of the landmark surface.

Within such a graph, all the true candidates are expected to be included in a *clique* (*i.e.*, a subset of nodes fully connected by edges), which indicates no conflict between them and thus justifies their co-occurrence. Therefore, an intuitive solution for locating true candidates is to find the maximum clique in the graph and harvest its member nodes. However, maximum clique finding is NP-complete and time-consuming even using approximation algorithms [6]. For the sake of efficiency, we design an approximation solution which relies on the confidence of each node namely visual phrase candidate. Starting from the most confident node, we iteratively select the most confident node that is connected to all the selected ones, until no more nodes can be added. In such a greedy manner, a set of most likely visual phrase occurrences are finally identified (as illustrated in Fig. 6 (d) and (e)), serving as 3D-to-2D matches for landmark recognition.

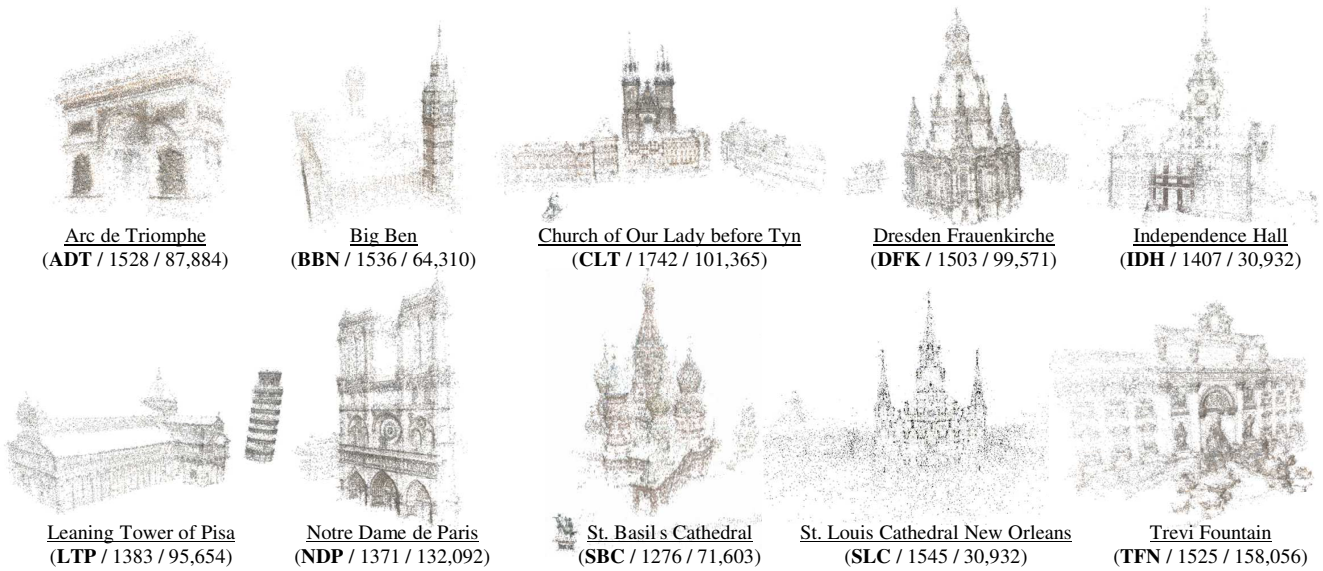


Figure 7. An overview of the landmarks and the reconstructed 3D point clouds. Each landmark is shown with the (abbreviated) name, the number of images registered to the point cloud, and the number of 3D points in the point cloud.

5. Evaluation

5.1. Experimental Settings

Dataset. To evaluate the effectiveness of the proposed method, we constructed a dataset consisting of images for 10 popular landmarks [2]. For each landmark, we collected several thousands of images from Flickr by issuing textual and geographical queries, and randomly sampled a set of images with known focal lengths (estimated from the EXIF tags) for 3D reconstruction. An overview of the dataset with the reconstructed 3D point clouds is shown in Fig. 7.

As the positive test set for each landmark, another collection of web images was crawled, manually checked, and diversified by (global feature based) clustering to form a compact set of 200 test images. To obtain negative test images for all the landmarks, we adopted the publicly available Oxford Buildings Dataset [13] consisting of 5062 Flickr images, because it contains both common noisy images and building images of particular Oxford landmarks, and thus is proper for evaluation of landmark recognition task.

Performance Metrics. Landmark recognition is essentially a classification problem, where a fundamental task is to determine whether a query image is relevant to a landmark [25]. Without loss of generality, in this paper we are concerned with such a binary decision task, and naturally adopt *precision* and *recall* as the performance measures. Precision is the proportion of positive images out of all images suggested by a solution, and recall is the proportion of identified ones out of all positive images in the ground truth. With different thresholds for binary decision, a precision-recall curve is commonly used to measure the overall performance. As image retrieval is out of the focus of this paper, the related performance metrics (*e.g.*, mAP) are not included in the evaluation.

Methods. To compare with the proposed 3D visual phrase based landmark recognition solution (abbreviated as **3DVP**), we first involved a bag-of-visual-words [18, 13] based solution **BoW** as baseline, which classifies each query image by searching nearest neighbors from landmark images registered to the corresponding 3D model, based on a vocabulary of 1M visual words. Then we extended the **BoW** implementation to include 2D visual phrases (**2DVP**), following the instructions in [23]. We also implemented a state-of-the-art approach, Point-to-Feature (**P2F**) Matching [11], which leverages 3D models for landmark recognition by matching 3D points to 2D features in query images. According to [11], for each landmark, a compact set of 3D points is selected from the 3D point cloud to cover each registered image at least $K = 100$ times; the P2F matching relies on SIFT ratio test with the threshold ratio λ set to 0.7.

Each of the four methods can be combined with geometric verification as post-processing. The **BoW** and **2DVP** methods filter nearest neighbor images via planar homography estimation. **P2F** adopts camera pose estimation (6-point DLT algorithm [7] with RANSAC) to filter putative point-to-feature matches. Our solution **3DVP** can also be combined with pose estimation as post-processing by obtaining putative matches from the detected visual phrases.

[2] The entire dataset (including image thumbnails, SIFT features, reconstructed 3D point clouds, and test cases) is publicly available at http://landmark3d.codeplex.com/.

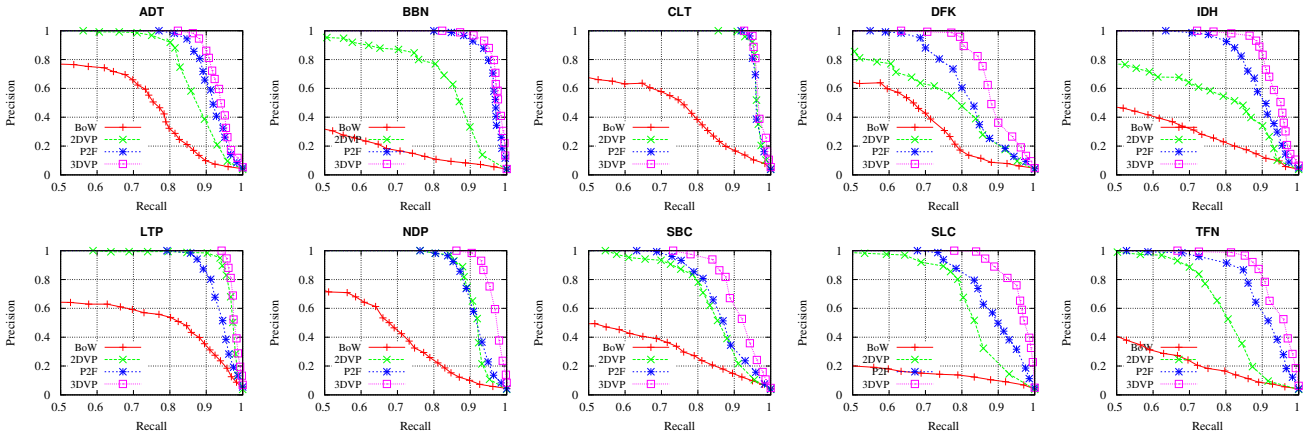


Figure 8. The precision-recall curves of different methods for recognizing 10 landmarks, *without* geometric verification.

5.2. Results

We compared both the *effectiveness* and *efficiency* of all the four methods for landmark recognition.

Effectiveness. For each landmark, we evaluated the recognition performance of all methods by matching test images with observed landmark images (for **BoW** and **2DVP**) or the 3D model (for **P2F** and **3DVP**). The resulting precision-recall curves are shown in Fig. 8, from which some observations can be drawn. First, **BoW** has the worst performance due to the lack of differentiation between landmarks and irrelevant objects in the same images, while **2DVP** gets significant improvement by considering spatial co-occurrence of visual words. Second, **P2F** is generally superior to **2DVP** because 3D points, compared with 2D descriptions, characterize the landmarks more accurately and ignore irrelevant objects. Third, **3DVP** outperforms other methods by leveraging 3D geometric structures of landmarks. The results also demonstrate that 3D visual phrases, compared with 2D visual phrases, are more repeatable and robust under projective transformations for landmarks observed from various viewpoints (*e.g.*, DFK, TFN).

We further evaluated the performance of all methods combined with geometric verification, which accepts a putative image if the estimated pose has sufficient inlier matches (*i.e.*, at least 8 inliers for a planar homography or 12 for a projection matrix [11]). According to the results listed in Table 1, geometric verification consistently leads to improvement in precision and decrease in recall. Particularly, homography estimation eliminates a large proportion of spurious landmark images suggested by 2D methods but some false positives still exist; while pose estimation, which is more strict to estimate a projective transformation, rejects all the false alarms. Compared with **P2F**, **3DVP** generally has less recall drops and keeps consistent advantages in recall after geometric verification, indicating that 3D visual phrases enable more reliable matching than 3D points due to the preserved geometric structures. As an interesting phenomenon, the recall drops extremely for landmarks with abundant repetitive structures (*e.g.*, LTP), which lead to spatial ambiguity in matching between local descriptions and therefore make putative matches difficult to agree on a consistent pose estimate; actually, this is a known challenge to the 3D reconstruction problem and is one of our future tasks. Note that as the evaluation is conducted on diversified test sets including sufficient tough cases, the resulting recall is lower than that reported in [11, 16], where the test sets are limited to the images that are empirically able to be registered to the 3D models.

Landmark		BoW		2DVP		P2F		3DVP	
		raw	+homo.	raw	+homo.	raw	+pose	raw	+pose
ADT	precision	0.09	0.24	0.39	0.96	0.59	1.00	0.99	1.00
	recall	0.91	0.88	0.89	0.72	0.91	0.80	0.86	0.82
BBN	precision	0.10	0.23	0.51	0.93	0.26	1.00	0.80	1.00
	recall	0.82	0.79	0.87	0.76	0.98	0.76	0.96	0.81
CLT	precision	0.17	0.23	0.35	0.86	0.89	1.00	0.99	1.00
	recall	0.90	0.88	0.97	0.95	0.94	0.92	0.94	0.93
DFK	precision	0.14	0.18	0.28	0.68	0.48	1.00	0.89	1.00
	recall	0.81	0.75	0.86	0.58	0.83	0.48	0.80	0.58
IDH	precision	0.08	0.20	0.27	0.66	0.36	1.00	0.97	1.00
	recall	0.95	0.87	0.92	0.69	0.93	0.56	0.86	0.64
LTP	precision	0.06	0.20	0.50	0.95	0.52	1.00	0.95	1.00
	recall	0.99	0.77	0.96	0.64	0.95	0.49	0.96	0.57
NDP	precision	0.08	0.26	0.36	0.85	0.98	1.00	1.00	1.00
	recall	0.92	0.86	0.92	0.87	0.80	0.66	0.86	0.71
SBC	precision	0.14	0.20	0.39	0.82	0.81	1.00	0.98	1.00
	recall	0.90	0.69	0.88	0.77	0.81	0.48	0.78	0.58
SLC	precision	0.11	0.21	0.52	0.90	0.22	1.00	0.83	1.00
	recall	0.86	0.65	0.83	0.73	0.96	0.51	0.90	0.62
TFN	precision	0.10	0.27	0.36	0.87	0.77	1.00	0.97	1.00
	recall	0.87	0.84	0.84	0.65	0.86	0.68	0.85	0.78

Table 1. Performance comparison of different landmark recognition methods *without vs. with* geometric verification.

Efficiency. As high efficiency is desirable in some landmark recognition scenarios, we also compared the speed of different methods. As reported in Table 2, the execution time consists of two parts, for (1) 2D-to-2D or 3D-to-2D matching and (2) geometric verification (RANSAC) which takes a maximum of 5,000 iterations for all methods.

Method	Matching	Geometric Verification		
	avg. time (s)	avg. time (s)	#verified img.	total time (s)
BoW	1.15 ± 0.27	0.16 ± 0.07	1839 ± 628	278.9 ± 118.2
2DVP	1.03 ± 0.24	0.14 ± 0.04	481 ± 111	58.9 ± 11.0
P2F	0.89 ± 0.28	0.09 ± 0.02	398 ± 232	37.9 ± 22.1
3DVP	0.86 ± 0.21	0.08 ± 0.01	189 ± 23	16.8 ± 4.3

Table 2. Efficiency comparison of different landmark recognition methods. All the statistics are averaged over the 10 landmarks.

The overall verification time of the four methods is in decreasing order, with a decreasing number of images that need verification. Particularly, the significant time reduction from **2DVP/3DVP** to their counterpart methods demonstrates the effectiveness of visual phrases for providing more geometrically reliable matches. To verify each image, the two 3D methods generally need less time than 2D methods because they tend to detect fewer putative matches with higher signal-to-noise ratio. With more confident matches than **P2F**, **3DVP** further reduces the time cost as well as the dependence on geometric verification.

As for the matching process, all the methods have comparable time cost. However, it should be noted that exhaustive 3D-to-2D matching is currently implemented for both **P2F** and **3DVP** methods because accuracy is our primary focus in this work. We also notice some acceleration strategies in the literature, including prioritized 3D-to-2D matching introduced in [11], and vocabulary-based prioritized search proposed in [16]. Such strategies could be naturally integrated with the proposed 3D visual phrases to improve the efficiency, which is in our recent plan.

6. Conclusion and Future Work

In this paper, 3D visual phrases have been proposed for landmark recognition. The basic idea is to incorporate spatial structure information to improve the discriminative ability of individual 3D points. In contrast to 2D visual phrases defined in 2D image planes, 3D visual phrases are derived from the physical space and explicitly characterize the 3D spatial structure of a landmark. Hence, they are inherently associated with descriptions that are highly robust to viewpoint changes. A complete solution has been proposed to *discover*, *describe*, and *detect* 3D visual phrases. Experiments on diverse data have shown promising performance of landmark recognition.

As a first attempt to leverage 3D visual phrases, the current solution still has room for improvement. There are several future directions. First, more geometric constraints are desired to afford to more relaxed point matching, as the missing of good correspondences still limits the recall. Second, we will accelerate the algorithms, especially the point matching step. And at last, we plan to extend 3D visual phrases to handle repetitive structures of landmarks.

References

[1] R. Arandjelovic and A. Zisserman. Efficient image retrieval for 3D structures. In *BMVC*, 2010.

[2] S. Arya and D. M. Mount. Approximate nearest neighbor queries in fixed dimensions. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 271–280, 1993.

[3] A. Broadhurst, T. Drummond, and R. Cipolla. A probabilistic framework for space carving. In *ICCV*, pages 388–393, 2001.

[4] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang. Spatial-bag-of-features. In *CVPR*, pages 3352–3359, 2010.

[5] O. Chum, M. Perdoch, and J. Matas. Geometric min-hashing: Finding a (thick) needle in a haystack. In *CVPR*, pages 17–24, 2009.

[6] U. Feige. Approximating maximum clique by removing subgraphs. *SIAM J. Discrete Math.*, 18(2):219–225, 2004.

[7] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[8] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *CVPR*, pages 2599–2606, 2009.

[9] J. Knopp, J. Sivic, and T. Pajdla. Avoiding confusing features in place recognition. In *ECCV (1)*, pages 748–761, 2010.

[10] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *Int. J. Comput. Vision*, 38(3):199–218, 2000.

[11] Y. Li, N. Snavely, and D. P. Huttenlocher. Location recognition using prioritized feature matching. In *ECCV (2)*, pages 791–804, 2010.

[12] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[13] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[14] R. Raguram, C. Wu, J.-M. Frahm, and S. Lazebnik. Modeling and recognition of landmark image collections using iconic scene graphs. *Int. J. Comput. Vision*, 95(3):213–239, 2011.

[15] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *International Journal of Computer Vision*, 66(3):231–259, 2006.

[16] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2D-to-3D matching. In *ICCV*, 2011.

[17] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *CVPR*, 2007.

[18] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.

[19] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Trans. Graph.*, 25(3):835–846, 2006.

[20] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *Int. J. Comput. Vision*, 80(2):189–210, 2008.

[21] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. Multicore bundle adjustment. In *CVPR*, pages 3057–3064, 2011.

[22] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In *CVPR*, 2007.

[23] S. Zhang, Q. Tian, G. Hua, Q. Huang, and W. Gao. Generating descriptive visual words and visual phrases for large-scale image applications. *IEEE Trans. Image Process.*, 20(9):2664–2677, 2011.

[24] Y. Zhang, Z. Jia, and T. Chen. Image retrieval with geometry-preserving visual phrases. In *CVPR*, pages 809–816, 2011.

[25] Y. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven. Tour the world: Building a web-scale landmark recognition engine. In *CVPR*, pages 1085–1092, 2009.

[26] C. L. Zitnick, J. Sun, R. Szeliski, and S. Winder. Object instance recognition using triplets of feature symbols. Technical Report MSR-TR-2007-53, Microsoft Research, 2007.