Distinguishing the Indistinguishable: Exploring Structural Ambiguities via Geodesic Context

Qingan Yan[1] Long Yang[1] Ling Zhang[1] Chunxia Xiao[21]
[1]School of Computer, Wuhan University, China
[2]State Key Lab of Software Engineering, Wuhan University, China
{yanqingan,yanglong,lingzhang,cxxiao}@whu.edu.cn

Abstract

A perennial problem in structure from motion (SfM) is visual ambiguity posed by repetitive structures. Recent disambiguating algorithms infer ambiguities mainly via explicit background context, thus face limitations in highly ambiguous scenes which are visually indistinguishable. Instead of analyzing local visual information, we propose a novel algorithm for SfM disambiguation that explores the global topology as encoded in photo collections. An important adaptation of this work is to approximate the available imagery using a manifold of viewpoints. We note that, while ambiguous images appear deceptively similar in appearance, they are actually located far apart on geodesics. We establish the manifold by adaptively identifying cameras with adjacent viewpoint, and detect ambiguities via a new measure, geodesic consistency. We demonstrate the accuracy and efficiency of the proposed approach on a range of complex ambiguity datasets, even including the challenging scenes without background conflicts.

(a) Background context

(b) Geodesic context

Figure 1. An example of visually indistinguishable scenes. (a) shows the symmetric structures of Temple of Heaven in Beijing. They are extremely similar in texture and contain little information in background for ambiguity inferring. Hence recent disambiguating methods would still lead to incorrect models for such scenes. (b) demonstrates that, rather than background context, the geodesic relationship explored by neighboring images provides a more meaningful way for ambiguity reasoning.

1. Introduction

Repetitive structures are widely existed in human world. When put them directly into 3D reconstruction, e.g., structure from motion (SfM), significant geometric errors would occur. Such reconstruction deficiency stems from the deceptive correspondence between ambiguous pictures. As in standard SfM pipelines [2, 11, 26], a pairwise feature matching [20] is usually applied first to establish visual connectivity across images. However, in the presence of repetitive structures, this step becomes very unreliable. Many similar but distinct features on different facades would be equally connected, which consequently misguide the following SfM process to register multiple instances onto a single surface, and give rise to hallucinating reconstruction models.

Distinguishing structural ambiguities is a challenging and vital task in SfM. Recent state-of-the-art methods [13, 16, 29] identify ambiguities relying highly on background context. That means, sufficient visual contradiction beside duplicate structures should exist within ambiguous images. Yet, for some scenes without noticeable background distinction, e.g., the Temple of Heaven shown in Fig. 1(a), the assumption will be violated.

The reason we, human observers, have the ability to distinguish ambiguities is likely because we can extract a global scene prior from the input collection, and then draw upon this knowledge to bridge the location gap between different views for decision making. In this paper we also intend to exploit this information. We note that the captured imagery of a scene often clusters along certain accessible paths in practice. Such assembled viewpoint collection reveals a high level knowledge on the global topology of input scene, that is, while ambiguous images appear deceptively similar in texture, they are actually located far apart according to the viewpoint variation (as illustrated in Fig. 1(b)).

We thus propose in this paper a novel geodesic-aware algorithm for visual ambiguity correction in SfM. Our basic idea is to characterize the available imagery using a manifold of viewpoints, and identify visual ambiguities through the intuition that a putative feature match should be not only visually connected but also geodetically consistent, which can be respectively encoded in two networks, *visibility network* and *path network*. We reason that the correspondences connected in visibility network but becoming unconnected according to the visual propagation along path network are geodetically inconsistent, i.e., ambiguous correspondences. Our algorithm is scalable and serves as a pre-process to the actual SfM reconstruction. We conduct 3D reconstruction on various challenging ambiguity datasets, and show correct registrations even in visually indistinguishable scenes.

In summary, we present three major contributions in this paper: (*i*) the idea of modeling the available imagery using a manifold of viewpoints for ambiguity correction in SfM applications, (*ii*) an embedding framework that organizes images geodetically onto manifolds in the presence of duplicate image content, and (*iii*) a new measurement, geodesic consistency, for adaptive ambiguity identification. Our code is available online at `https://github.com/yanqingan/SfM_Disambiguation`.

2. Related Work

Symmetric and duplicate structure has recently earned great interest in graphics and vision community. Such pattern provides an informative prior for applications, like image completion [15], monocular modeling [17, 25, 31], bundle readjustment [9] and scene stitching [8]. On the other hand, repetitive structures can also contribute to visual ambiguities in feature matching, which are disastrous to SfM. While recent matching systems [7, 19, 32, 33] have made significantly progress on efficiency and accuracy, they are still incapable of distinguishing ambiguous features. In this section, we revisit several kinds of related approaches that aim at mitigating the effect of structural ambiguity.

The first kind of work are based on geometric reasoning. Zach *et al*. [36] infer structural ambiguities by validating *loop consistency* over match graph. They reason that the cumulation of associate transforms between an image pair in a loop should be the identity. Any cycle involving obvious loop closure inconsistency indicates the emergence of incorrect registrations. However, this criteria limits the effectiveness of this approach over larger loops, as the accumulated errors in transform calculation would become non-ignorable. Ceylan *et al*. [6] present another method based on the idea of loop constraint. They first detect repetitive elements in each image via a user-marked pattern, then perform a graph-based optimization to obtain global consistent repetition results. This method makes significant improvements over Zach *et al*. [36] but specifies in only regular repetitions appearing on planar facades, thus can not handle rotational symmetries as found on domes, churches, etc.

Another mechanism for structural ambiguity is to explore background context. Zach *et al*. [35] introduce the concept of *missing correspondences*, where the main idea is to analyze the co-occurrence of feature correspondences among image triplets. If a third image loses a large portion of matches shared by the other two, then this view is more likely to be mismatched. Yet, this metric is also prone to rejecting many positive image pairs with large viewpoint variation punitively. Roberts *et al*. [22] improve the criteria by integrating it with the image timestamp cue into an expectation maximization (EM) framework and estimating misregistrations iteratively. Such temporal information makes their method more accurate, whereas also limits its usage in unordered images. Jiang *et al*. [16] introduce a novel objective function that evaluates global missing correspondences upon the entire scene instead of image triplets. They argue that a correct 3D reconstruction should associate to the minimal missing of reprojected 3D points in images. This assumption is reasonable, however, it also fails on unordered photo collections.

Therefore, more recently, Wilson and Snavely [29] extend the idea of missing correspondences to large-scale Internet collections. They validate if the neighboring observations within one image are also visible to other pictures and adopt *bipartite local clustering coefficient* (blcc) to quantify such consistency. This algorithm is quite scalable, but not well suited for small-scale datasets. In addition, it easily leads to over-segmentations, as all detected bad *tracks* are directly discarded. Heinly *et al*. [13] introduce a useful post-processing framework for ambiguity correction by analyzing reprojected geometry conflicts between images. They first obtain an initial 3D model via SfM, then detect and mitigate mis-registration errors in SfM by comparing the 2D projections of distinctive 3D structures. Taking inspiration from this work, Heinly *et al*. [14] design another post-processing approach by efficiently analyzing the co-occurrence of 3D points across images using *local clustering coefficient* (lcc). These two methods are functioned in many challenging scenes, however, incur some computational cost, as a reconstructed 3D model is required. Moreover, for the scene without explicit background distinction, they would also lead to poor performance.

In this work, we explore a totally different property from recent existing disambiguation methods. Our method investigates the geodesic relationship among photo collections and makes no assumption on sequence information or background context. This enables our method to tackle a range of challenging photo collections, where recent approaches

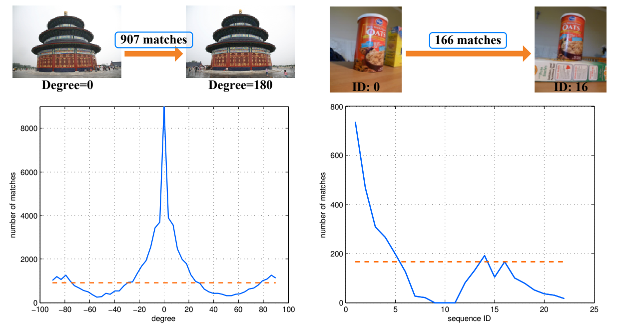


Figure 2. The statistics of feature matches according to viewpoint changes. Although these images look the same, many matches are still missing and can only be matched by neighboring images with similar viewpoint.

either fail or work poorly. Another advantage of our approach is the scalability. Our method serves as a pre-process to the incremental SfM and works automatically and efficiently even on large-scale Internet datasets. Furthermore, our method does not delete a bad track directly, instead, we separate it into multiple distinct individuals. So it enables us to produce unbroken models.

3. Modeling Ambiguity on Manifolds

As in standard structure from motion (SfM) setups [26], we assume that a collection of images $\mathbf{I} = \{I_1, ..., I_n\}$ about a desired scene is available, associated with a set of feature matches acquired by image matching [20] and geometric verification [10]. More specifically, the relationship between images and correspondences can be expressed as a bipartite graph $V = (\mathbf{I}, \mathbf{T}, \mathbf{L})$, called *visibility network* [29]. It has nodes respectively for images $\mathbf{I}$ and tracks $\mathbf{T}$, where a *track* T_f refers to a sequence of local features capturing the same physical point f within different image planes, and an edge $(I_i, T_f) \in \mathbf{L}$ exists if the spatial point represented by track T_f is visible in I_i. We denote a *visual connection* as an edge pair (L_{if}, L_{jf}) linked by the same track.

SfM operates on tracks for image registration and camera attachment. Normally, a track should correspond to a unique 3D point in the physical world, however, in the presence of duplicate structures, the feature matching step is prone to blending multiple 3D points into a single track. Our objective is thus to validate the plausibility of visual connections in V and decompose those blended tracks. We note that even if duplicate structures look the same in appearance, they are actually situated in different geographical positions, i.e., there are location conflicts between them. Fig. 3 shows a briefly illustration of our idea.

3.1. Path Network Representation

In order to tractably estimate the location conflict without known camera poses, we convert this hard wide-baseline problem into many easier small-baseline pieces by investigating the viewpoint trajectory as encoded topologically in the *path network*. Formally, the network $G = (\mathbf{I}, \mathbf{E})$ has nodes for every image $I_i \in \mathbf{I}$, and edges $(I_i, I_j) \in \mathbf{E}$ linking image pairs that have geodetically neighboring viewpoint, e.g., sideward movement.

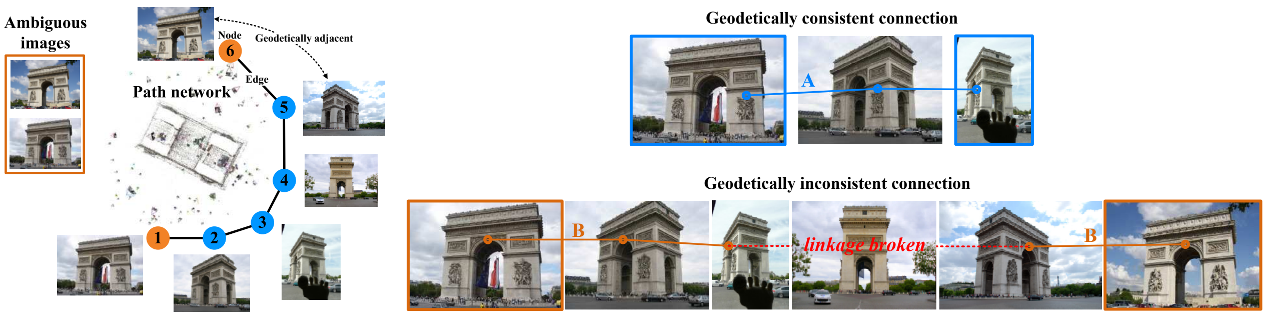
This is based on three useful observations. (1) In practice, photographers always snap a scene along certain accessible streets, which can be described as a sequence of paths within path network. (2) For the purpose of 3D reconstruction, the input scene is often over-pictured from different viewpoints. Such abundant visual overlaps can serve as the adjacent sampling nodes along network paths. (3) Ambiguous structures are just texturally similar but never exactly the same, as the example shown in Fig. 2. That implies, the geodetically neighbors usually contain more useful information, which is meaningful for network construction, than duplicate copies.

The many geodetic neighbors for each image provide us a high level knowledge on the global topology of 3D scene for ambiguity reasoning rather than a single image content. We also define that a *geodesic path* $\mathbf{P}_{ij} = \{E_{i*}, ..., E_{*j}\}$ refers to a sequence of connected edges linking image I_i and I_j. Note that such path $\mathbf{P}_{ij}$ actually reveals the virtual camera trajectory from viewpoint I_i to I_j. Duplicate images with similar appearance always consist of a distant geodesic path in the network as shown in Fig. 3(a).

3.2. Geodesic Consistency

Yet our scheme is not to calculate geodesic distances (shortest paths [5, 27]) directly between each image pair in the network, as it is difficult to decide the exact threshold that corresponds to the emergence of visual ambiguity. Many ambiguity-free images with wide viewpoint variation may also contribute to large distance values. We propose instead a new metric, *geodesic consistency*, to quantitatively measure the contradiction. We note that if two images are mismatched, the geodesic paths between them are either blocked or visually disjointed; it is unable to propagate what they see from one node to another according to the viewpoint variation in path network, even though they are visually connected in visibility network.

To illustrate this concept more clearly, we refer to the example in Fig. 3. In Fig. 3(a), we show a path network consists of six images (with IDs in colored circles) from dataset **Arc de Triomphe**. The images directly linked by an edge (the black solid line) are geodetically adjacent. Fig. 3(b) shows two tracks A and B (as plotted in blue and orange respectively), which offer us the visibility of three 3D points (B corresponds to two points), and two geodesic paths within the network: from image 1 to 3 and from image 1 to 6. In order to validate the plausibility of visual connection associated to track A between image 1 and 3. We



(a) Path network (b) Geodesic consistency

Figure 3. A simple illustration of our geodesic-aware disambiguating strategy. (a) shows a path network consisting of six images in **Arc de Triomphe**. Note that while the two ambiguous images look similar in appearance, they share a long geodesic path. (b) shows two tracks and two geodesic paths. The visual connection corresponding to track A between image 1 and 3 is geodetically consistent, as all intermediate nodes – image 2, in this path also observe this track. However, the connection between image 1 and 6 for track B is inconsistent; the track is lost in intermediate image 4 along its geodesic path.

check if all intermediate nodes along its geodesic path also observe this track. Since track A is visible to image 2, the connection thus satisfies the criteria of geodesic consistency and is considered plausible. In contrast, the connection between image 1 and 6 is implausible, where its geodesic path consists of six images. While image $1, 2, 3$ all observe track B, in image 4, the track is lost. This disconnection in visibility propagation provides the evidence of scene change and indicates that image 1 and 6 actually correspond to distinct 3D points, i.e., geodetically inconsistent.

Therefore, the geodesic consistency criteria requires that the correct pairwise connections ought to be transmissible based on their geodesic paths in network G. More specifically, let L_{ip} denotes an edge in V between node I_i and T_p. We define that *a visual connection* (L_{ip}, L_{jp}) *associated to track* T_p *is geodetically consistent, only if existing a feasible geodesic path* $\mathbf{P}_{ij} = \{E_{ik}, ..., E_{kj}\} \subset \mathbf{E}$ *links image* I_i *and* I_j*, and each intermediate node* I_k *along this path observes the track, i.e.,* $L_{kp} \in \mathbf{L}$*; otherwise, the connection is ambiguity-affected.* We formulate this measure as below:

$$H(L_{ip}, L_{jp}) = \begin{cases} 1 & \mathbf{P}_{ij} \neq \emptyset, \\ -1 & \text{otherwise.} \end{cases} \quad (1)$$

3.3. Objective

We now have a metric that capably determines ambiguous connections. However, we do not mean to remove tracks that contain inconsistent connections directly, instead we find a way to reuse them. Our objective can thus be expressed simply in the form of:

$$Q_L = \sum_{T'_p \in \mathbf{T}'} H(L'_{ip}, L'_{jp}) - \sum_{T_p \in \mathbf{T}} H(L_{ip}, L_{jp}), \quad (2)$$

which needs to be maximized. We take V as the only input. $V' = (\mathbf{I}, \mathbf{T}', \mathbf{L}')$ is a new disambiguated visibility network we intend to achieve.

Term $\sum H(\cdot)$ evaluates the total quality of edge pairs in V according to the geodesic consistency criteria. If there are undesired connections, they would contribute to a negative increase to the evaluation. By dividing confusing tracks into distinct ones, we could block the negative contribution originating from inconsistent visual connections and get a larger Q_L. In comparison, incorrect splitting of plausible tracks would cause a decrease of positive edge pairs. Thus, intuitively, the global maximum of Q_L should correspond to a correct visibility network. If the dataset is ambiguity-free, all visual connections contribute to a positive value, so $Q_L = 0$ and $V' = V$.

4. Disambiguating Algorithm

Our algorithm accordingly has two main steps: (1) construct a path network, then (2) revise ambiguous tracks based on the analysis of geodesic consistency. We next describe each of them in turn.

4.1. Network Construction

A main technical problem we face is the establishment of path network. It is challenging to acquire desirable geodesic relations from images without known cameras poses or geotags, particularly in the presence of ambiguous image content. Recent image embedding methods, like knn-based methods [4, 27] or training-based methods [12, 28] do not take ambiguity into consideration, so they are not valid alternatives in our cases. Although [23] use a ranking-based method for sideways image retrieval, it is also insufficiently robust for most ambiguity datasets. To overcome this challenging problem, we propose a useful sample-and-grow strategy, which exploits both explicit and implicit unique points within images for neighborhood inference.

Scene sampling phase Normally, a 3D scene can be decomposed into two categories of points: *confusing points*

$\mathbf{D}$ which contribute to ambiguities, and *unique points* $\mathbf{U}$ that do not cause visual ambiguity. Such unique points are meaningful information for our network construction. Furthermore, unique points also include two groups: explicit unique points (e.g., salient background conflicts, such as explored in [13, 16, 29]) and implicit unique points (corresponding to small-scale texture variation). As shown in Fig. 2, even if ambiguous photos look extremely similar, there are still many features that can only be matched by their geodetically neighbors. This suggests that, beyond explicit background distinction, there are also many implicit unique points in foreground can be used.

In order to identify unique points (both the explicits and implicits), we summarize the scene by selecting a set of iconic images. In particular, we require that the selected samples $\mathbf{C} \subset \mathbf{I}$ should satisfy two properties: (*i*) *completeness*, i.e., covering the scene as complete as possible, and (*ii*) *distinctiveness*, which means the iconic images ought to be sufficiently distinctive from one another in appearance.

Such iconic views provide an overview of the input scene. Additionally, we note that due to the existence of unique points within foreground and the requirement of scene completeness, the representative images corresponding to repetitive structures, e.g., the front and back gate of **Arc de Triomphe**, would be also respectively selected. By intersecting these iconic images, we can then get the confusing points that contribute to geometric ambiguities; on the other hand, the remaining points consequently are unique points. To formulate, let $\mathbf{T}_i$ denotes the tracks observed by image I_i. Given the iconic set, we approximate the overall points of this scene by $\mathbf{T}_A = \bigcup_{I_i \in \mathbf{C}} \mathbf{T}_i$, where $\mathbf{T}_A \subseteq \mathbf{T}$. The confusing points are therefore expressed as the tracks that are observed by more than one iconic image, where $\mathbf{D} = \bigcup_{I_i, I_j \in \mathbf{C}} \mathbf{T}_i \cap \mathbf{T}_j$, and accordingly the unique points can then be computed via $\mathbf{U} = \mathbf{T}_A - \mathbf{D}$.

To obtain the iconic images adaptively, we formulate above properties into two objective terms. Term *completeness* can be expressed as $|\bigcup_{I_i \in \mathbf{C}} \mathbf{T}_i|$, which describes the number of tracks that are covered by set $\mathbf{C}$. We expect this term to be as large as possible in order to ensure a good coverage of the input scene. In the meanwhile, term *distinctiveness* is quantified in the form of $|\bigcup_{I_i, I_j \in \mathbf{C}} \mathbf{T}_i \cap \mathbf{T}_j|$, which measures the collision of tracks contained in the set. This term prevents us choosing redundant candidates from similar viewpoint. Consequently, our sampling process is then equivalent to maximize the following quality function:

$$R(\mathbf{C}) = |\bigcup_{I_i \in \mathbf{C}} \mathbf{T}_i| - \alpha |\bigcup_{I_i, I_j \in \mathbf{C}} \mathbf{T}_i \cap \mathbf{T}_j|, \tag{3}$$

where $\alpha > 0$ controls the effect of distinctiveness term.

We solve the optimization problem in an efficient greedy manner, similar to [24]. This scheme begins with $\mathbf{C} = \emptyset$ and $R(\mathbf{C}) = 0$. At each iteration, we calculate $\Delta = R(\mathbf{C} \cup I_i) - R(\mathbf{C})$ for each image I_i in the photo collection and choose the view I_*, for which its Δ_* is maximal. If $\Delta_* \geq 0$, we then add this view into $\mathbf{C}$ as a new iconic image. The iteration proceeds until no view in the collection can make $\Delta_* \geq 0$.

Path growth phase Given the selected iconic images and unique points, this phase involves the computation of linkages between iconic views and the other images according to unique points. In this respect, the path network G can be looked upon as a bipartite graph with nodes respectively are iconic images and non-iconic images.

According to the calculation of unique points $\mathbf{U}$, they would be uniquely distributed in each iconic image; there are no common points between any two of them. The unique points $\mathbf{U}_i$ contained in each iconic image I_i actually indicate the scene that should be visible in neighboring non-iconic images. We therefore define that an image pair is *geodetically adjacent* only if they share common unique points. Formally, for each non-iconic image I_j, we add a direct edge to the network between I_j and any iconic image I_i satisfying $|\mathbf{U}_j \cap \mathbf{U}_i| > \epsilon$, where ϵ is a small positive constant for the consideration of noise tracks. We use $\epsilon = 5$ in our experiments.

Note that the acquired path network is meaningful; the selected iconic images form the basic anchors of the input scene, whereas non-iconic images serve as paths relating these isolate points. This embedding scheme is efficient and performs well on our experimental datasets.

4.2. Track Regeneration

With the available path network, our remaining computation is then to obtain a disambiguated visibility network that maximizes the objective in Eq. 2. Rather than exhaustively validate geodesic consistency on each visual connection in V, we take an efficient propagation approach.

Initially, the disambiguated visibility network V' is empty. For each image I_i, we investigate its direct neighbors in G. If the track T_f in V, shared by I_i and one of its neighbors I_j, is already observed by image I_i (or I_j) in V' in the form of track T'_f, then we associate the other view I_j (or I_i) also to this existed track T'_f; otherwise, we create a new track to represent the connection between I_i and I_j in V'. This scheme makes the visibility gradually propagate from one image to its neighbors in path network until all images and their direct neighbors have been processed. It creates new tracks only when it is necessary, so guarantees the optimum to Eq. 2. Additionally, there is no need to explicitly compute geodesic consistency, as these directly connected neighbors are geodetically consistent.

In practical implementation, this procedure can be regarded as a step of re-computing tracks based on the neighborhood in path network and can be done efficiently by traveling the network in breadth-first order.

5. Experiments

In this section, we evaluate the performance of our proposed algorithm on a wide variety of photo collections that are associated with visual ambiguities. They are common examples in our daily life, ranged from small-scale laboratory objects to large-scale urban structures. Table 1 lists a detailed summary of these datasets.

There is only one parameter α used in our method. This makes our approach a viable option for general use. We found that the value $\alpha = 0.1$ is sufficient to produce satisfactory results in our experiments. We implement the algorithm in C++ and test it on a machine of 3.30GHz Xeon quad-core CPU, along with 32GB memory.

We first validate the robustness of our method on a set of benchmark datasets for correct SfM reconstruction. Dataset **Oats** [22] is obtained by sampling around an indoor object using a handheld camera. Thus it is relatively small in scale and has uniform image resolution and illumination condition. It is interesting to note that there are no duplicate objects in this scene; it is the same object placed in different places. So it contains little implicit unique points in foreground, but the existence of massive explicit unique points in background supports our inference. Additionally, we found in experiment that our scene sampling algorithm in Sec 4.1 is able to identify the entirety of confusing points in this scene, while accompanying about 38% negative selections. Yet fortunately, the over-identification of a certain amount of confusing points, in some cases, would not cause much trouble, as long as there still are enough unique points remaining in each image to indicate geodesic inference.

In contrast, unstructured photo collections **Arc de Triomphe**, **Alexander Nevsky Cathedral** and **Berliner Dom**, acquired from [13], are much larger in scale and contain images with various resolutions and illuminations. They all exhibit a closure of a landmark architecture, however, due to the existence of repeated structures, some parts are misplaced. These datasets contain both a high quantity of explicit and implicit unique points, which makes our method easily rectify ambiguous tracks and yield correct 3D models. For **Arc de Triomphe**, we successfully recover its two facades in opposite directions while keeping them unbroken. For **Alexander Nevsky Cathedral**, our method is able to prune the hallucinating dome stemming from duplicate structures, and correct the mis-registration along the river of **Berliner Dom**.

Moreover, We also test our algorithm on separate models, such as **Radcliffe Camera** [13] and **Sacre Coeur** [29]. Like [13], our method succeeds to identify the two ambiguous facades of **Radcliffe Camera**. Yet due to the missing of available images linking these two facades, the disambiguated model is also divided into two parts. **Sacre Coeur** suffers from the same problem, but more challenging. There are many structures causing ambiguity, such as the sideway facades, extra towers and domes. Similar to [29], our algorithm achieves the four parts of this model: the front and two sides of the building, and an overview towards Paris. The results on these datasets are shown in Fig. 4.

Table 1. Performance statistics of our algorithm on different photo collections. From top to bottom, the datasets respectively are **Sacre Coeur**, **Berliner Dom**, **Alexander Nevsky Cathedral**, **Arc de Triomphe**, **Radcliffe Camera**, **Temple of Heaven**, **Cup**, **Building** and **Oats**. N_{img} and N_{pt} indicate the number of input cameras and reconstructed 3D points respectively.

Dataset	N_{img}	N_{pt}	Time		
			Ours	[29]	[13]
SC	4,530	590,268	51.4 m	6.1 m	–
BD	1,618	241,422	11.9 m	3.2 m	11.8 h
ANC	448	92,820	2.3 m	36 s	33.4 m
AdT	434	92,055	2.2 m	21 s	39.7 m
RC	282	77,623	1.2 m	28 s	–
ToH	145	127,752	2.0 m	18 s	26.7 m
Cup	64	8,810	27 s	3 s	2.5 m
Bd	47	14,895	36 s	2 s	2.0 m
Oats	23	8,585	10 s	1 s	45 s

To evaluate the specialty, besides benchmark collections, we also test our method on several challenging datasets, where recent disambiguating systems work poorly or fail. Dataset **Cup** [16] shows a single cup with duplicate textures on opposite surfaces. The only available background context is the cup handle (as exploited in [16]), whereas it is hard to detect via super-pixel segmentation in [13]. Dataset **Building** exhibits a series of highly repetitive facades on a building. These pictures are taken along a straight street and contain rare distinctive structures. Moreover, we also test our algorithm on the difficult challenge of **Temple of Heaven** in [16] (serving as one of their limitations). This rotationally symmetric architecture looks nearly the same from any direction, while exhibiting negligible features in background. A commonality of these examples is the difficulty in discrimination by making use of missing correspondences [29] or conflicting observations [13]. In contrast, our algorithm exploits not only the explicit unique points in background for ambiguity reasoning, but also implicit unique points within foreground. We correctly recover the camera trajectory and symmetric geometry of these scenes. Fig. 5 shows our disambiguating results.

In Table 1, we record the detailed performance statistics of our system, including the number of input images and reconstructed points, and the runtime (including disambiguation and I/O process) of each compared method. Our algorithm is much more efficient than [13] and has a wider application scope as compared to [13, 16, 29]. Since it serves as a pre-process for SfM, we do not require the availability of camera poses and 3D point locations in advance. We test [29] and our algorithm on only one core, whereas [13]

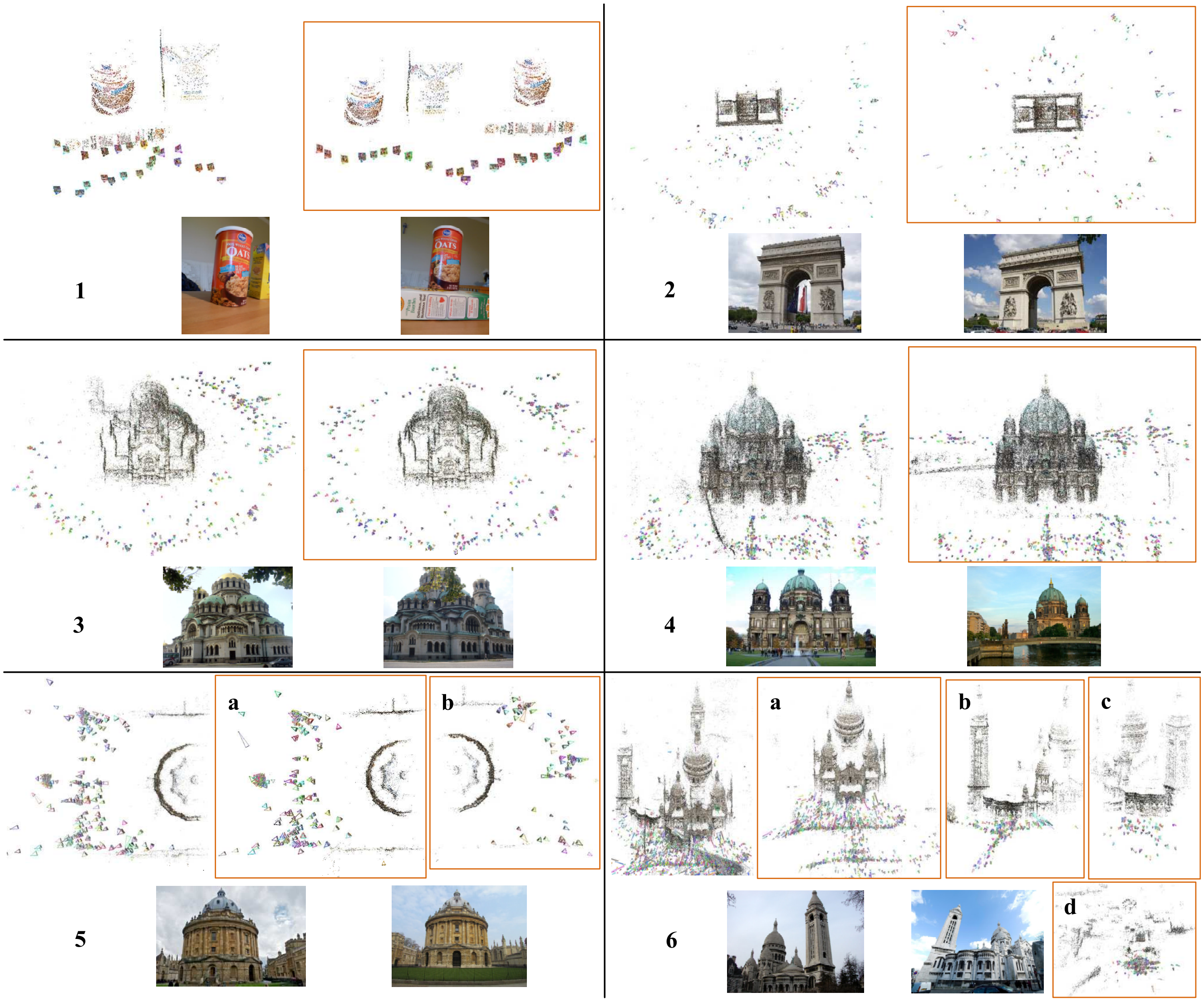


Figure 4. Disambiguation results of our method on benchmark datasets. From 1 to 6: **Oats**, **Arc de Triomphe**, **Alexander Nevsky Cathedral**, **Berliner Dom**, **Radcliffe Camera** and **Sacre Coeur**. The left pictures show the results produced by VisualSFM [30]. The right images marked in orange are the results acquired by our proposed algorithm.

is performed on 4 threads.

Comparison with [29] For further comparison, we run the Matlab code from [29] on datasets in our experiments. This method also serves as a pre-process to standard SfM reconstruction and takes the visibility network as input. However, it additionally requires an FOV (field of view) file. The main advantage of this method is its scalability. It can be seen from our statistics on runtime performance in Table. 1. This algorithm is extremely fast as compared to [13] and ours. However, it suffers from a big limitation on accuracy. While **Sacre Coeur** is correctly separated, many other datasets are over-segmented, such as **Radcliffe Camera** and **Berliner Dom**. The main reason attributes to this phenomenon is the punitive removal of bad tracks. In addition, it also has failures on **Oats** and visually indistinguishable datasets, due to the limited images for blcc validation and the lack of background information.

Comparison with [13] To compare with this work, we also test their Matlab code and use the thread pool set to be 4. This method is much more robust than [29] and performs well on most datasets in our experiments due to the existence of sufficient background context. However, it fails on visually indistinguishable datasets as well. For **Cup**, VisualSFM provides a roughly correct point cloud but with rare background conflicts for further improvement, so this method outputs the input with geometry unchanged. For **Building** and **Temple of Heaven**, it also fails to identify duplicate structures due to the lack of useful conflicting ob-

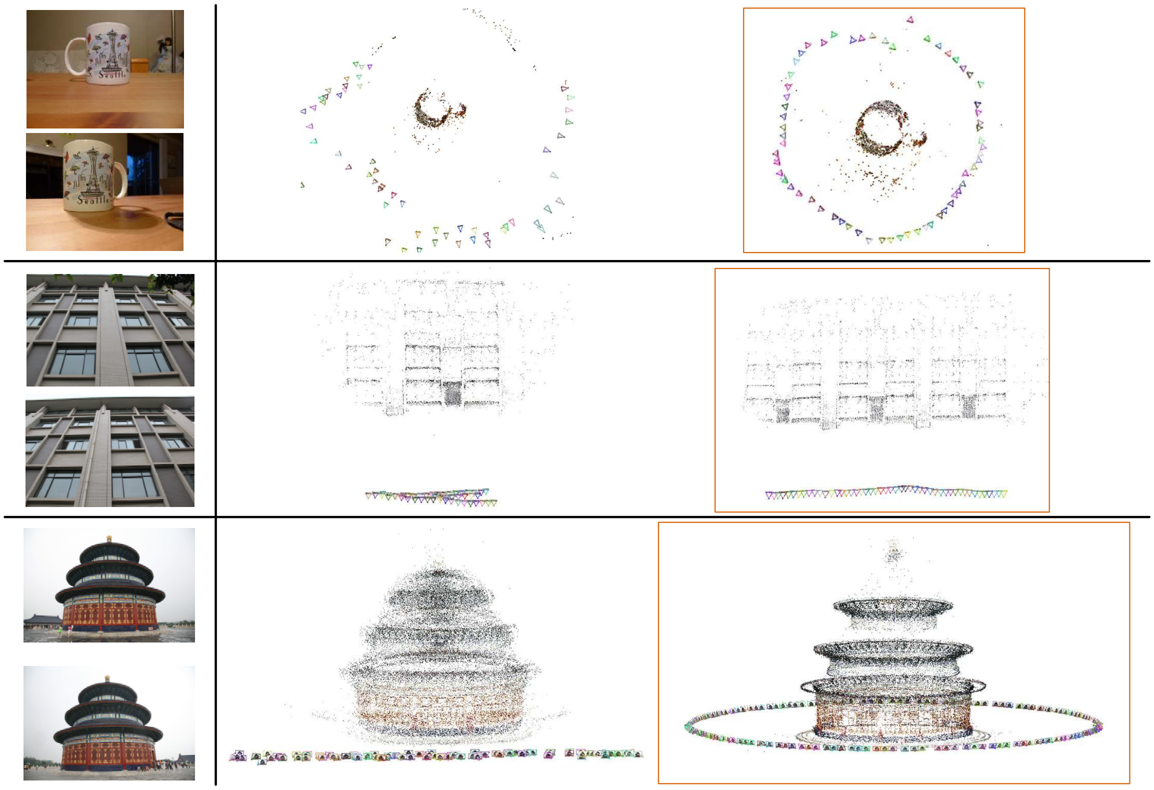


Figure 5. Results on several challenging datasets with visually indistinguishable repetitions, which respectively are **Cup**, **Building** and **Temple of Heaven**. The second column shows the results of [30]. The third column exhibits the 3D models generated by our method.

servations. Another deficiency of the approach is its high computational cost. It requires an initial SfM model as input and relies on SLIC [1] to detect super-pixels in each image. So in order to disambiguate **Berliner Dom**, it spends us more than 11 hours and 20GB memory spaces. Additionally, we always encounter parallelization errors in Matlab when test **Radcliffe Camera** on different machines, and suffer from an overflow on **Sacre Coeur**.

Limitations Although we have demonstrated the effectiveness of our method on diverse datasets, we also note several limitations. First, deriving path network from images is a challenging problem. In order to produce satisfactory results, we implicitly assume that there are sufficient viewpoint overlaps (usually less than 60 degrees) between images. We have visualized the curve of matches according to viewpoint variation in Fig. 2. The lack of reasonable viewpoint overlaps around duplicate instances, such as two photo clusters taken at widely different scales about one identical building, may affect the accuracy of geodesic inference in our path network construction. Second, the greedy search in scene sampling phase in Sec. 4.1 could get stuck at a local minimum. For instance, consider the reconstruction result (left facade) of **Arc de Triomphe** in Fig. 4. Due to the over-selection of iconic images, some positive tracks that do not cause ambiguity are considered as confusing points and eliminated in path network construction. This leads several images to remain isolated in path network and could not be linked in track regeneration step.

6. Conclusion

In this paper, we have presented a new geodesic-aware method to remedy SfM ambiguity caused by repetitive structures, which can be considered as a valid complement to background context. We note that the input imagery approximates a manifold of viewpoints and ambiguous views fall apart on this manifold. We propose a useful framework to infer geodesic relationship from images in the presence of ambiguity, and a meaningful measure to quantify ambiguity. We show that this method is accurate and efficient and can handle a variety of challenging examples even without informative background context.

The path network provides an intuitive way for scene understanding. Thus in the future, it might be fruitful to extend the geodesic prior to SLAM [18, 21, 34] for loop-closure detection, and SfM scene analysis [3, 8].

Acknowledgments This work was partly supported by the NSFC (No.61472288, 61672390), NCET (NCET-13-0441), and the State Key Lab of Software Engineering (SKLSE-2015-A-05). Chunxia Xiao is the corresponding author.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[2] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. In *ICCV*, pages 72–79, 2009.

[3] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, pages 1534–1543, 2016.

[4] H. Averbuch-Elor and D. Cohen-Or. Ringit: Ring-ordering casual photos of a temporal event. *ACM Transactions on Graphics*, 34(3):33, 2015.

[5] J. Carreira, A. Kar, S. Tulsiani, and J. Malik. Virtual view networks for object reconstruction. In *CVPR*, pages 2937–2946, 2015.

[6] D. Ceylan, N. J. Mitra, Y. Zheng, and M. Pauly. Coupled structure-from-motion and 3d symmetry detection for urban facades. *ACM Transactions on Graphics*, 33(1):2, 2014.

[7] J. Cheng, C. Leng, J. Wu, H. Cui, and H. Lu. Fast and accurate image matching with cascade hashing for 3d reconstruction. In *CVPR*, pages 1–8, 2014.

[8] A. Cohen, T. Sattler, and M. Pollefeys. Merging the unmatchable: Stitching visually disconnected sfm models. In *ICCV*, pages 2129–2137, 2015.

[9] A. Cohen, C. Zach, S. N. Sinha, and M. Pollefeys. Discovering and exploiting 3d symmetries in structure from motion. In *CVPR*, pages 1514–1521, 2012.

[10] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[11] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, et al. Building rome on a cloudless day. In *ECCV*, pages 368–381. 2010.

[12] C. Hegde, A. C. Sankaranarayanan, and R. G. Baraniuk. Learning manifolds in the wild. *Preprint, July*, 2012.

[13] J. Heinly, E. Dunn, and J.-M. Frahm. Correcting for duplicate scene structure in sparse 3d reconstruction. In *ECCV*, pages 780–795. 2014.

[14] J. Heinly, E. Dunn, and J.-M. Frahm. Recovering correct reconstructions from indistinguishable geometry. In *3DV*, volume 1, pages 377–384, 2014.

[15] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf. Image completion using planar structure guidance. *ACM Transactions on Graphics*, 33(4):129, 2014.

[16] N. Jiang, P. Tan, and L.-F. Cheong. Seeing double without confusion: Structure-from-motion in highly ambiguous scenes. In *CVPR*, pages 1458–1465, 2012.

[17] K. Köser, C. Zach, and M. Pollefeys. Dense 3d reconstruction of symmetric scenes from a single image. In *Joint Pattern Recognition Symposium*, pages 266–275, 2011.

[18] G. H. Lee and M. Pollefeys. Unsupervised learning of threshold for geometric verification in visual-based loop-closure. In *ICRA*, pages 1510–1516, 2014.

[19] W.-Y. Lin, S. Liu, N. Jiang, M. N. Do, P. Tan, and J. Lu. Repmatch: Robust feature matching and pose for reconstructing modern cities. In *ECCV*, pages 562–579, 2016.

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[21] M. Pollefeys, D. Nistér, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S.-J. Kim, P. Merrell, et al. Detailed real-time urban 3d reconstruction from video. *International Journal of Computer Vision*, 78(2-3):143–167, 2008.

[22] R. Roberts, S. N. Sinha, R. Szeliski, and D. Steedly. Structure from motion for scenes with large duplicate structures. In *CVPR*, pages 3137–3144, 2011.

[23] J. L. Schönberger, F. Radenović, O. Chum, and J.-M. Frahm. From single image query to detailed 3d reconstruction. In *CVPR*, pages 5126–5134, 2015.

[24] I. Simon, N. Snavely, and S. M. Seitz. Scene summarization for online image collections. In *ICCV*, pages 1–8, 2007.

[25] S. N. Sinha, K. Ramnath, and R. Szeliski. Detecting and reconstructing 3d mirror symmetric objects. In *ECCV*, pages 586–600. 2012.

[26] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. *ACM Transactions on Graphics*, 25(3):835–846, 2006.

[27] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[28] M. Torki and A. Elgammal. Putting local features on a manifold. In *CVPR*, pages 1743–1750, 2010.

[29] K. Wilson and N. Snavely. Network principles for sfm: Disambiguating repeated structures with local context. In *CVPR*, pages 513–520, 2013.

[30] C. Wu. Visualsfm: A visual structure from motion system. `http://homes.cs.washington.edu/~ccwu/vsfm`, 2011.

[31] C. Wu, J.-M. Frahm, and M. Pollefeys. Repetition-based dense single-view reconstruction. In *CVPR*, pages 3113–3120, 2011.

[32] Q. Yan, Z. Xu, and C. Xiao. Fast feature-oriented visual connection for large image collections. *Computer Graphics Forum*, 33(7):339–348, 2014.

[33] Q. Yan, L. Yang, C. Liang, H. Liu, R. Hu, and C. Xiao. Geometrically based linear iterative clustering for quantitative feature correspondence. *Computer Graphics Forum*, 35(7):1–10, 2016.

[34] L. Yang, Q. Yan, Y. Fu, and C. Xiao. Surface reconstruction via fusing sparse-sequence of depth images. *IEEE Transactions on Visualization and Computer Graphics*, 2017.

[35] C. Zach, A. Irschara, and H. Bischof. What can missing correspondences tell us about 3d structure and motion? In *CVPR*, pages 1–8, 2008.

[36] C. Zach, M. Klopschitz, and M. Pollefeys. Disambiguating visual relations using loop constraints. In *CVPR*, pages 1426–1433, 2010.