# Progressive Large Scale-Invariant Image Matching In Scale Space

Lei Zhou lzhouai@cse.ust.hk
Siyu Zhu szhu@cse.ust.hk
Tianwei Shen tshenaa@cse.ust.hk
Jinglu Wang jwangae@cse.ust.hk
Tian Fang* tianft@cse.ust.hk
Long Quan quan@cse.ust.hk

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology

## Abstract

*The power of modern image matching approaches is still fundamentally limited by the abrupt scale changes in images. In this paper, we propose a scale-invariant image matching approach to tackling the very large scale variation of views. Drawing inspiration from the scale space theory, we start with encoding the image's scale space into a compact multi-scale representation. Then, rather than trying to find the exact feature matches all in one step, we propose a progressive two-stage approach. First, we determine the related scale levels in scale space, enclosing the inlier feature correspondences, based on an optimal and exhaustive matching in a limited scale space. Second, we produce both the image similarity measurement and feature correspondences simultaneously after restricting matching between the related scale levels in a robust way. The matching performance has been intensively evaluated on vision tasks including image retrieval, feature matching and Structure-from-Motion (SfM). The successful integration of the challenging fusion of high aerial and low ground-level views with significant scale differences manifests the superiority of the proposed approach.*

## 1. Introduction

The past few years have witnessed the growing application of image-based 3D reconstruction due to the convenience of image capturing, the progress of scalable reconstruction algorithms [8, 52, 50, 17, 11, 12, 18, 33] and the advance of 3D toolkit [45, 44, 48]. As the first yet nontrivial step of standard SfM pipeline [1, 10, 39, 52], image matching generally first selects similar image pairs from unstructured image sets, then establishes exact feature correspondences between selected pairs for accurate camera registration and structure recovery. However, modern matching approaches would get trapped when handling large differences of view scales, *e.g.*, the fusion case of street-to-aerial urban reconstruction [38, 4].

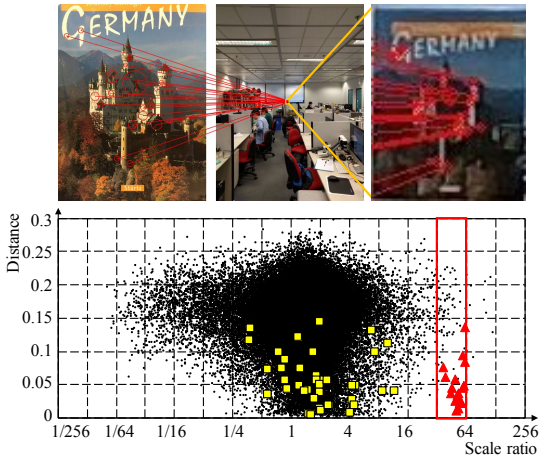


Figure 1: The top row visualizes the SIFT matches found by our approach between images holding scale ratio around 55. The black dots in below chart represent 69640 putative matches, characterized by feature scale ratio and descriptor distance. The yellow squares denotes the false positive correspondences found by modern SIFT match strategies [23, 46, 43, 9]. Conversely, our approach first determines the related scale levels in scale space that enclose the feature correspondences. Then image matching is restricted between the related scale levels. The red box encloses the 90 feature pairs existing between related scale levels, including 20 inlier correspondences found eventually denoted by red triangles.

The difficulties can be summarized in two aspects. On one hand, devising scale-invariant similarity metric to identify whether two images share visual overlap is hard. On the other hand, establishing exact feature correspondences albeit scale variations is harder. Concretely, image matching across large scale differences has to include small-scale features to establish correspondences. However, overwhelm-

*Tian Fang is with Shenzhen Zhuke Innovation Technology since 2017.

ing noisy and ambiguous feature pairs are thus introduced and the matching process becomes much more prone to mismatches. A matching example of two images holding scale ratio around 55 is shown in Figure 1. Massive putative SIFT [23] matches are generated by nearest neighbor search as denoted by the dense black dots in the below chart. The false positive matches denoted by yellow squares are given after passing the mutual best [43] and ratio test [23] filtering as well as RANSAC verification [9]. The distance-based filtering schemes [43, 23] could not overcome the noise because many outlier matches even hold much smaller distances than inlier ones due to the limitation of descriptor discrimination and local patch ambiguity. RANSAC [9] is neither robust in this case because there exist sufficient outlier matches that could follow certain geometric agreement.

Since it is intractable to undertake matching directly, we draw inspiration from the scale space theory [21, 20] and propose a novel scale-invariant image matching approach. As the preprocessing step, we divide the image scale space into multiple scale levels and encode it into a compact multi-scale representation. Based on the representation, two progressive matching steps are taken. The first step is the exhaustive yet efficient scale level matching in limited scale space. A scale level matching map that describes matching responses between scale levels is thus obtained. Further analysis of the map is performed to determine the related scale levels in scale space that roughly enclose the inlier feature correspondences. The second step is the scale-aware image matching which restricts matching between the related scale levels and finally gives both the similarity measurement and feature correspondences robustly at the same time. As shown in Figure 1, the first step remarkably reduces the excessive match pairs to a small quantity of candidates enclosed by the red box. Hence, the rare inlier correspondences denoted by red triangles are found in the second step under protection against disturbing unrelated feature pairs.

Our contributions are two-fold. 1) The progressive image matching approach restricts the correspondence search of query features within limited related scale space and thus boosts the accuracy and robustness of feature matching under drastic scale variations. 2) Since the matching scope is narrowed down in scale space, the matching efficiency gets improved.

## 2. Related Work

To extend the matching ability of handling large scale variations, abundant previous work has been proposed for both image similarity measurement and feature matching.

For similarity measurement, Bag-of-Features (BoF) model is widely used in image retrieval context [40, 31, 32]. However, it is generally very hard for flat BoF model to distinguish images with large scale differences from the negative ones due to lack of overlap. To enhance the capability of retrieving different scale images, the query expansion technique [7] is equipped by [25, 26, 37]. The query is repeatedly expanded and re-issued to retrieve spatially-consistent images with slight scale differences by absorbing new information from the growing scale range. But the drifted expansion of the query has strong dependence on image database in which smooth scale transition is required. Besides, Jegou *et al*. [13] first proposed and then Li *et al*. [19] adopted the constraint of weak geometric consistency for more robust similarity scoring. Arandjelovic *et al*. [2] uses MultiVLAD heuristically to generate VLAD representations for sub-images. But all of these would be out of order under large scale variations where overwhelming noise occurs.

In terms of feature matching, the key is to solve the feature correspondence problem. Generally, putative matches are first found by exhaustive search or approximate search [23, 29, 30, 42, 6], combined with some effective filtering strategies like mutual best [43] and ratio test [23]. Finally, robust statistical methods like RANSAC [9] follow to apply geometric constraints and reject outliers. In the milestone paper [23], Lowe has proposed the high-performance SIFT matching scheme. A number of extensions, including SURF [3] for speed, ASIFT [28] for full affine invariance, are proposed. However, all the matching approaches [23, 30, 42, 6, 3, 28] are devised to tackle feature matching at similar scales fundamentally. In some special cases, Shan *et al*. [38] has to rely on geo-information to apply view-dependent matching between ground and aerial images.

Complementary to all the work above, we dissect the matching problem from the perspective of scale space [21, 20] in this paper and seek to make both the similarity metric and feature matching scale-invariant and independent of any other auxiliary knowledge.

## 3. Scale-Invariant Image Matching

### 3.1. Overview

As observed by the literature [13, 19], two images perceiving the same scene from similar viewpoints follow the *scale consistency*: the scale ratios of inlier feature correspondences are centralized around the value which we term *image scale ratio*. Illustrative examples can be found in Figure 2. The image scale ratio indicates that the matched images are merely related in limited scale space.

Motivated by this, we propose a scale-invariant matching approach organized as follows:

- **Image scale space encoding.** We divide the scale space of an image into multiple scale levels evenly. Then the BoF-based encoding strategy is applied to represent an image compactly at multiple scale levels for subsequent progressive matching.

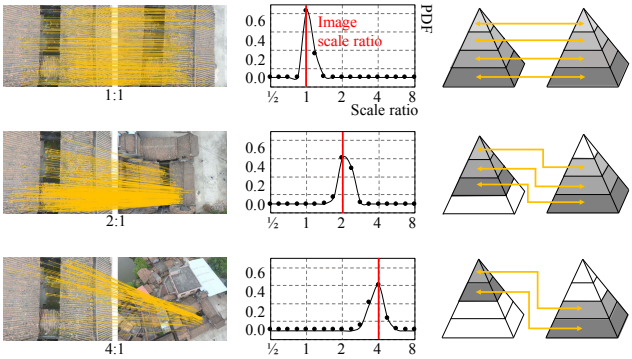


Figure 2: Illustration of three match pairs with image scale ratio around 1, 2, and 4. The left column visualizes the feature matches. The middle column is the PDF of scale ratios of the inlier feature matches. Following the scale consistency [13, 19], the scale ratios are quite centralized. Deduced from the scale consistency, inlier correspondences are supposed to link between the scale levels in scale space with consistent discrepancy as shown in the right column.

- **Scale level matching.** Two sets of scale levels are matched exhaustively in limited scale space and a scale level matching map denoting the matching responses are generated. Then by imposing the scale consistency, we convert the problem of discovering scale level relations to detecting the line pattern in the matching map.
- **Scale-aware image matching.** After obtaining the scale level relations, image matching including both similarity measurement and detailed feature matching is restricted between the related scale levels.

### 3.2. Image Scale Space Encoding

Effective representation of an image is an open problem. The basic requirements are *distinctiveness* for distinguishing and *repeatability* for matching. With respect to current representation methods [5, 34, 15, 31, 16, 49], repeatability could not be preserved under large scale variations. In other words, the representations of matched images with large scale difference are distant in representation space. To achieve scale-invariant repeatability, we seek to encode the scale space [21] of an image into its representation. The scale space theory [21, 20] considers image representation at all scales simultaneously by successively suppressing fine details with Gaussian smooth. Therefore, we combine the compact representation method [31] and the idea of scale-space representation [21] to encode the scale space in a compact and comparable way.

First, we discretize the image's scale space into multiple scale levels with equal logarithmic range and group features into the scale levels by their scales. Without loss of generality, we use SIFT features [23] below to elaborate on the proposed algorithm. Scale space division in SIFT paradigm can be naturally taken as layers of DoG pyramid:

$$\bigcup_{l\in\{1,...,L\}} \{f|\ell(f)=l\} \triangleq \bigcup_{l\in\{1,...,L\}} \mathcal{L}_l, \qquad (1)$$

where $\ell(f)$ gives the scale level indexed by the parameter $l \in \{1, ..., L\}$ from which feature $f$ is extracted. Since nearby scale levels are separated by a constant multiplicative factor, the feature scales tend to ascend and the feature number tends to descend exponentially when the scale level parameter increases.

Next, the BoF framework is adopted here to vectorize scale levels, as it is the de-facto standard way to encode local features [31, 13, 32, 16, 49]. Pre-trained from 128-dimensional SIFT descriptor vectors in corpus, an overcomplete codebook is first constructed including a total of $K$ visual words, *i.e.*, $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_K] \in \mathbb{R}^{128\times K}$. Then, the descriptor vector $\mathbf{x}_n$ is assigned to the nearest visual word $\mathbf{v}_k$, resulting in a unit vector $\mathbf{u}_n \in \mathbb{R}^K$ satisfying $\mathbf{V}\cdot\mathbf{u}_n = \mathbf{v}_k$. After that, by summing up all the unit vectors, the scale level $\mathcal{L}_l$ is represented by the distribution histogram of the visual words:

$$\mathbf{p}_l = \mathbf{w} \odot \Big(\sum_{n=1}^{N} \mathbb{1}_{\mathcal{L}_l}(\mathbf{x}_n)\cdot\mathbf{u}_n\Big), \qquad (2)$$

where $N$ denotes the image feature number and $\mathbb{1}_{\mathcal{L}_l}(.)$ is the indicator function of subset $\mathcal{L}_l$. The Hadamard product $\odot$ with weight vector $\mathbf{w} \in \mathbb{R}^K$ assigns weights to each visual word based on the tf-idf scoring scheme [31].

Finally, for an image comprising $L$ scale levels, it can be represented by a $K \times L$ matrix by stacking all its scale levels after normalization:

$$\mathbf{P} = [\mathbf{p}_1, ..., \mathbf{p}_L]. \qquad (3)$$

The multi-scale image representation explicitly encodes all scale levels simultaneously and thus holds more powerful describing capability in face of scale variations.

### 3.3. Scale Level Matching

After obtaining the multi-scale image representation, we aim to discover how the scale levels of an image pair are related. A scale level matching map is first generated by exhaustive scale level matching. Then the image scale ratio is approximately determined by analyzing the matching map while the scale consistency is imposed.

**Scale Level Matching Map** Since no prior information is available, exhaustive scale level matching is performed. The matching response, which can also be interpreted as similarity, of two scale levels $\mathbf{p}$ and $\mathbf{q}$ is computed by the normalized L2 distance of their representations, *i.e.*,

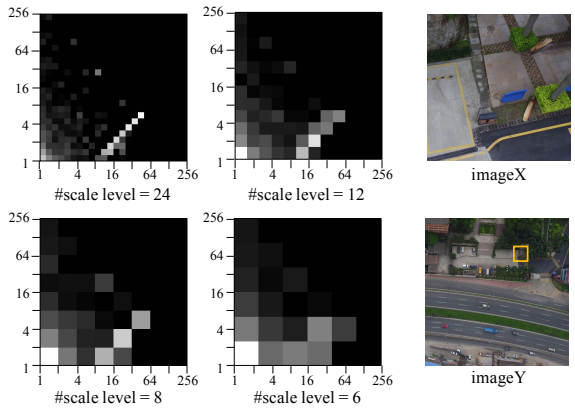


Figure 3: Scale level matching maps of images X and Y in the right column with scale ratio around 10. The overlap region is marked in yellow box. The scale range of 1 to 256 is divided into different numbers of scale levels. The higher brightness of each square reflects the stronger matching response between corresponding scale levels. Following the scale consistency deduction in Equation 5, line patterns in anti-diagonal direction are observed indicating constant discrepancy between related scale levels.

$s(\mathbf{p}, \mathbf{q}) = \frac{\mathbf{p}^\top}{||\mathbf{p}||_2} \cdot \frac{\mathbf{q}}{||\mathbf{q}||_2}$. Then the scale level matching between two image representations $\mathbf{P}$ and $\mathbf{Q}$ comprising $L$ scale levels can be manipulated simply by matrix multiplication:

$$\mathbf{M} = \mathbf{P}^\top \cdot \mathbf{Q}. \tag{4}$$

The resultant *scale level matching map* $\mathbf{M}$ is an $L \times L$ matrix and its element $\mathbf{M}(i, j) = \mathbf{p}_i^\top \cdot \mathbf{q}_j$ equals the matching response of i-th scale level to j-th scale level of two images. Intuitively, the exhaustive scale level matching seems costly. In fact, however, the matrix multiplication is quite efficient because of the sparsity. The representation matrix of an image with $N$ features has at most $N$ nonzero elements due to the hard assignment of descriptors to visual words. Therefore, for two images with respectively $N_1$ and $N_2$ features, the scale level matching takes no more than $\min(N_1, N_2)$ float multiplications.

**Scale Ratio Determination** Because inlier feature correspondences share consistent scale ratios [13, 19], it can be deduced that the matches only link between scale levels with consistent discrepancy. Concretely, let $\tau$ denote the scale factor separating nearby scale levels and $\rho$ denote the image scale ratio. Then the discrepancy of related scale levels $\mathcal{L}_{l_a}$ and $\mathcal{L}_{l_b}$ should approximately satisfies the relationship formulated as

$$l_a - l_b \approx log_\tau \rho. \tag{5}$$

For better elaboration, an illustrative matching map of two images with scale ratio around 10 is displayed in the first picture in Figure 3. The scale range from 1 to 256 times is evenly divided in logarithmic scale into 24 scale levels, so the nearby scale levels are separated by the factor $2^{1/3}$. Most elements in the map assume zero response due to scarcity of features in top scale levels and high irrelevance between most scale levels. However, strong responses can be observed along an anti-diagonal line which corresponds to the constant discrepancy of scale levels around 9. Besides, the discrepancy and the image scale ratio satisfy Equation 5 precisely, which justifies the deduction based on scale consistency [13, 19] formulated in Equation 5.

Therefore, to determine the scale ratio is equivalent to detecting the anti-diagonal line pattern in the matching map. Specifically, we search along each anti-diagonal line and average all its elements. Then the averaged value is taken as the matching response with respect to the scale level discrepancy $\Delta l \in \{1 - L, ..., L - 1\}$:

$$r(\Delta l) = \begin{cases} \frac{1}{L - \Delta l} \sum_{l=1}^{L-\Delta l} \mathbf{M}(l, l + \Delta l) & \Delta l \geq 0 \\ \frac{1}{L + \Delta l} \sum_{l=1-\Delta l}^{L} \mathbf{M}(l, l + \Delta l) & \Delta l < 0. \end{cases} \tag{6}$$

Finally, the desired scale level discrepancy is determined by the discrepancy at which maximum matching response is assumed:

$$\Delta l^* = \arg\max_{\Delta l} r(\Delta l). \tag{7}$$

And the image scale ratio can be approximately computed by substitute the discrepancy $\Delta l^*$ into Equation 5.

**Discussion** The previous works, like WGC [13] and PGM [19], also try to estimate the image scale ratio for filtering out spurious matches. However, these Hough-Voting-based techniques are quite vulnerable to noisy feature correspondences caused by too large image scale ratio. The reason is that the votes of scale ratio bins are obtained by accumulating the number of feature pairs [19] or weights of the common visual words to which the feature pairs are assigned [13]. As a result, the votes exhibit a strong bias towards the scale ratio near 1, because images generally have most of their features at the lowest scale level and thus share most putative feature correspondences at the lowest scale level as well. Hence, the bias would mislead the Hough Voting methods to the noisy and plausible estimation of the scale ratio. Conversely, our method greatly suppresses noise by two steps: First, divide the scale space into sliced scale levels; Second, obtain the matching response for each scale ratio hypothesis as the normalized similarity between the related scale levels. The effect of scale space division in suppressing noise is illustrated in Figure 3. If the scale space is divided into less scale levels, matching responses of unrelated scale levels increase because excessive noisy feature pairs are introduced. When it comes to only one scale level, it degenerates to the general case [23, 31, 13, 19] where the full feature sets are matched directly. On the other hand,

if the number of scale levels gets too large, a scale level containing too few features can not be characterized effectively and there is an increasing cost in terms of memory consumption. As a good trade-off, we follow SIFT [23] to divide each octave (separated by 2 times) of scale space into 3 scale levels in our experiments.

### 3.4. Scale-Aware Image Matching

Image matching gets easy after we are aware of the relations between image scale levels.

For similarity measurement, a scale-invariant similarity metric is devised by averaging the matching responses of the related scale levels. Formally the similarity between two image representations $\mathbf{P}$ and $\mathbf{Q}$ is measured by

$$S(\mathbf{P}, \mathbf{Q}) = \max_{\Delta l} r(\Delta l) = r(\Delta l^*), \tag{8}$$

following Equation 6 and 7. After filtering out unrelated information, the metric is able to provide scale-invariant similarity measurement robustly for image pairs.

In terms of feature matching, the correspondence problem is simplified into inter-level matching between related scale levels. For a query feature $f^q \in \mathcal{L}_l^q$, the search scope is narrowed down to the related scale level $\mathcal{L}^s_{l+\Delta l^*}$ of search image. Considering the quantization error when discretizing image scale space, we conservatively extend the search scope by including neighboring scale levels, *i.e.* $\mathcal{L}^s_{l+\Delta l^* \pm 1}$, if they have non-zero matching responses against the query scale level $\mathcal{L}_l^q$. Within the two feature subsets, prevalent matching framework can be applied in a more robust way involving exact or approximate search [23, 29, 30, 42, 6] for putative matches, combined with filtering strategies [23, 43] and RANSAC [9] for outlier rejection.

The advantage brought by the inter-level matching is that the number of reliable inlier correspondences gets raised. This is because, after shrinking the search range in scale space, more inlier correspondences survive thanks to the *distinctiveness relaxation*, *i.e.*, the constraint of feature distinctiveness will be relaxed merely inside the search scope. Specifically, filtering strategies like ratio test [23] and mutual best [43] are widely applied in feature matching. They are very effective in rejecting many false matches but also some proportion of correct matches as expense. However, if the match scope is restricted to two small subsets, feature pairs just need to be discriminative within the subsets rather than the whole feature sets. Hence, more inlier feature pairs could be preserved as practically verified by experiments in Section 4.2.

### 3.5. Complexity Analysis

The matching efficiency is critical for match-intensive tasks like SfM [39]. Assuming matching between two images each containing $N$ features, the complexity of brute-force (BF) search is $O(N^2)$. Approximate nearest neighbor (ANN) search [23, 29, 30] generally reduces the complexity to $O(N \cdot logN)$ while sacrificing the precision to some degree. Cheng *et al*. [6] further reduces the complexity through cascading hashing technique. On the contrary, the overhead of our matching approach lies in two aspects: the scale level matching and the inter-level feature matching. Given the pre-trained vocabulary tree, the expenditure of descriptor quantization and BoF scoring by scale level matching is linear to the feature number. With respect to the inter-level feature matching, the complexity by brute-force search is turned into linearity to the number of involved candidate pairs between related scale levels. And the larger the scale difference is, the lower the complexity would be. The comparison of runtime with respect to image scale ratio using BF search is displayed in Figure 4. Certainly, the ANN search [23, 29, 30] and hashing search [42, 6] can also be equipped for acceleration.

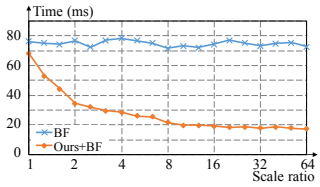


Figure 4: Time of pairwise matching with respect to the image scale ratio between around 50k×50k SIFT sets implemented by GPU. Our approach keeps on improving the efficiency of brute-force (BF) search as the scale ratio increases.

## 4. Experiments

In this section, we would like to evaluate the performance of the proposed scale-invariant matching approach. Both the image similarity metric and the feature matching are evaluated on both the benchmark datasets and our own datasets with large scale variations. Moreover, a large ground-aerial SfM experiment is conducted based on the collaboration of similarity metric and feature matching.

### 4.1. Similarity Metric Evaluation

The scalable retrieval technique in [31] is used as the baseline here, following the process of building a hierarchical vocabulary tree with branch factor 8 and maximum leaf size 16 from the full image corpus and quantizing SIFT descriptors [23] into leaf nodes. The proposed scale-invariant similarity metric can be readily integrated into the implementation. The only variant is that we replace the original scoring scheme with the proposed scale-invariant similarity metric in Equation 8.

Firstly, the proposed similarity metric is evaluated on the benchmark datasets Oxford [32] and Holidays [13]. The statistical results are summarized in Table 1. Although very rare and limited scale variations occur in the two datasets,

| Dataset | Oxford | | Holidays | |
|---|---|---|---|---|
| | mAP | Time (s) | mAP | Time (s) |
| Baseline | 0.515 | **840.9** | 0.607 | **697.1** |
| Baseline + Ours | **0.562** | 859.6 | **0.646** | 780.6 |

Table 1: Statistical image retrieval results on benchmarks Oxford [32] and Holidays [13] achieved by the baseline [31] and the baseline+our similarity metric.

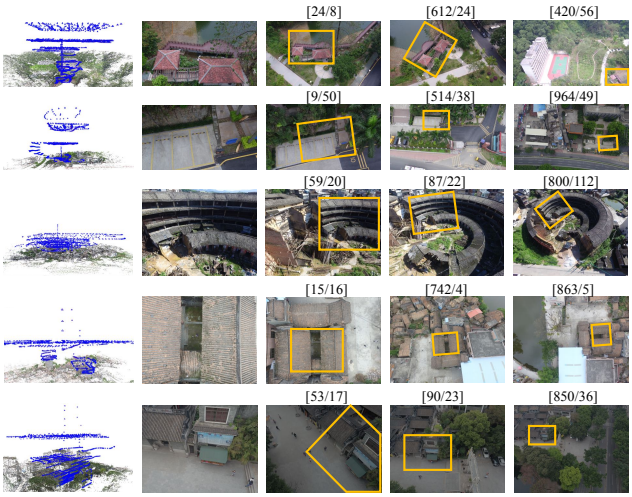


Figure 5: Our multi-scale retrieval dataset including 5 scenes. In each row, camera poses are first displayed, then one of the 5 queries in each scene followed by 3 sample database images at different scales. Common areas shared with the query are marked by yellow bounding boxes. The two numbers in square brackets above each image are the ranks given by the baseline [31] and baseline+our similarity metric respectively.

| Teniques | None | WGC | HE | HE+ WGC | MA | PGM | PGM+ HE+MA |
|---|---|---|---|---|---|---|---|
| Before | 0.397 | 0.440 | 0.490 | 0.524 | 0.455 | 0.451 | 0.554 |
| After | **0.672** | **0.679** | **0.712** | **0.722** | **0.695** | **0.680** | **0.734** |

Table 2: Comparison of mAPs achieved by methods that combine different techniques, WGC [13], HE [13], PGM [19], MA [14], with the baseline approach [31]. After our similarity metric is applied in cooperation with these techniques, improvements in mAP by a significant margin are observed.

moderate improvements in mean average precision (mAP) are still observed without significant increase of running time.

To further validate the retrieval performance in the presence of scale variations, we additionally collect a *Multi-scale* image set comprising 7349 high resolution images and 5 different scenes, with maximum image scale ratio reaching 10. Without loss of generality, we choose 5 closely-captured images as queries for each scene. To obtain the ground-truth match labels, we rely on the constructed 3D model of each scene and project the mesh triangles to database images to determine if they share overlap with the queries. Camera motions and sample images of the dataset are shown in Figure 5. The ranks of typical multi-scale database images given by the baseline approach equipped with or without our similarity metric are also given. The baseline method suffers from degeneracy when scale differences increase. But our scale-invariant similarity metric effectively preserves the high ranks of relevant images despite scale variations. In the literature, a set of techniques have been proposed for the image retrieval context, such as weak geometric consistency (WGC) [13], Hamming Embedding (HE) [13], pairwise geometric matching (PGM) [19], Multiple Assignment (MA) [14] and so on. As the proposed approach only revises the similarity metric, it is actually complementary to almost all of these techniques [13, 19, 14]. Therefore, we compare the performance of the retrieval techniques before and after using our similarity metric, as shown in Table 2. The mAPs all get boosted by a significant margin after our metric is integrated. It demonstrates the superiority of the proposed approach in tackling large scale variations.

| Dataset | Method | #Matches | | | | | Average time (ms) |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | |
| Zoom+ Rotation | SIFT | 2309 | 1773 | 603 | 374 | 112 | 59.16 |
| | SIFT+Ours | **2506** | **1918** | **716** | **490** | **178** | **33.26** |
| | Improve (%) | 8.5 | 8.2 | 18.7 | 31.0 | 58.9 | 43.8 |
| Viewpoint | SIFT | 1270 | 369 | 64 | 0 | 0 | 27.89 |
| | SIFT+Ours | **1433** | **499** | **91** | 0 | 0 | **23.99** |
| | Improve (%) | 12.8 | 35.2 | 42.2 | N/A | N/A | 14.0 |

Table 3: Comparative matching results on Mikolajczyk benchmark [24]. The statistics show the number of correct matches achieved by exhaustive SIFT matcher [23] and our scale-invariant matcher under increasing distortion of zoom plus rotation and viewpoint change from pair 1 to 5. The match number and speed get improved in all cases by our method. Both matchers fail under too large viewpoint change in pair 4 and 5 of *Viewpoint* set.

## 4.2. Feature Matching Evaluation

Feature matching performance is evaluated by the number of inlier feature correspondences found eventually. The proposed scale-invariant matcher is implemented following the procedure: First, the scale-invariant feature descriptors are fed into a pre-trained vocabulary tree [31] and scale level matching is conducted to determine the related scale levels. Then putative matches between related scale levels are found, combined with robust filtering strategies including mutual best [43], ratio test [23] and RANSAC [9] to finally produce geometrically-consistent matches.

In this part, we apply our matcher to SIFT features [23] and compare it with the exhaustive SIFT matcher [23] on the standard Mikolajczyk dataset [24]. The dataset is composed of two image sets emphasizing scale and view point changes respectively. Five image pairs (reference image vs. images 1 to 5) are included in each set with increasing distortion. Both matchers are implemented based on the SiftGPU library [46] for acceleration. As shown in Table 3, our matcher outperforms SIFT matcher in all cases in term of match number thanks to the *distinctiveness relaxation* as analyzed in Section 3.4. And the improvement is monotone

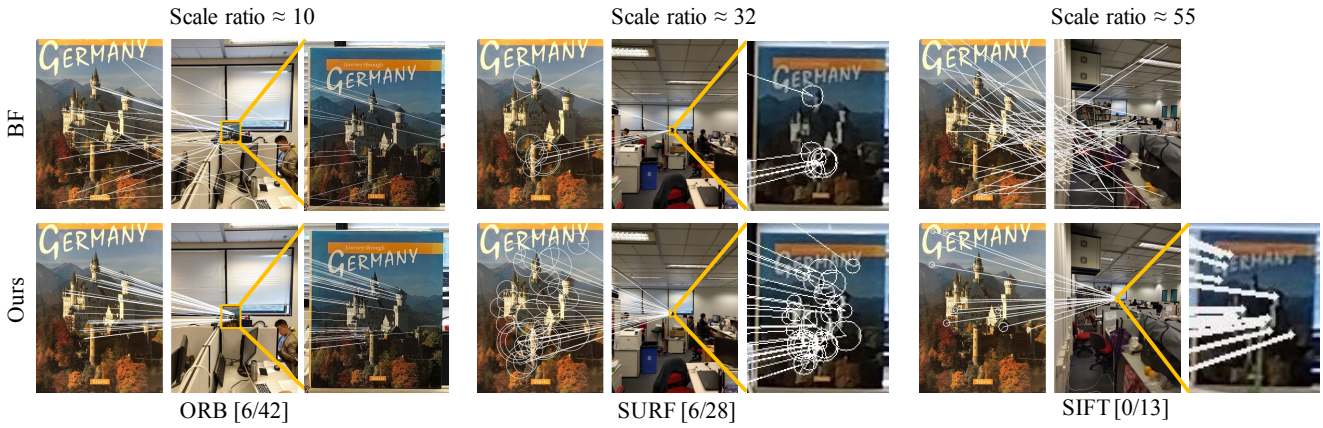


Figure 6: Visualization of matching results using ORB [35], SURF [3] and SIFT [23] features by brute-force (BF) matcher and our scale-invariant matcher. The two numbers in below square brackets denote the correct match numbers achieved by two methods respectively.

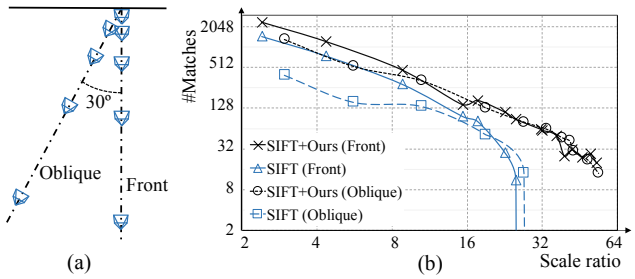


Figure 7: (a) Capture pattern of *book-cover* image set. (b) Number of correct matches with regard to image scale ratio in logarithm scale achieved by SIFT matcher and ours. SIFT matching fails when the image scale ratio exceeds 25 while our methods shows its superiority in both match number and robustness to scale variations even when the image scale ratio approaches 60.

increasing as the distortion gets larger. Besides, less time is consumed. As a common problem with SIFT matching, our matcher is also limited by too large viewpoint changes in *Viewpoint* set.

To further explore the robustness of our matching approach to scale variations, we collect a *book-cover* image set composed of two series of photographs acquired from two viewpoint angles 0° and 30° with increasing zoom as illustrated in Figure 7(a). The 28 images have the same size of 3024×4032 pixels. The closet-captured photograph of a book cover is used as reference image and is matched against other 27 ones. The number of correct SIFT matches with regard to varying image scale ratio between 2 and 64 is summarized in Figure 7(b). Our scale-invariant matcher achieves larger numbers of feature matches at both frontal and oblique views and all scales. Exhaustive SIFT matcher [23] totally fails when scale ratio goes beyond 25 while our method successfully matches all 27 image pairs even though the scale ratio approaches 60. Furthermore, our scale-invariant matching approach is compared with the state-of-the-art matching methods MODS [27] and KVLD [22] on this image set. For fair comparison, our approach shares the same SIFT detection process as KVLD and the same feature description process as MODS. As reported in Table 4, our matching approach outperforms the other two in terms of the match number.

| Scale ratio | 8 | 16 | 32 | 36 | 42 | 48 | 55 |
|---|---|---|---|---|---|---|---|
| MODS | 129 | 99 | 19 | 0 | 0 | 0 | 0 |
| KVLD | 411 | 120 | 44 | 16 | 3 | 0 | 0 |
| Ours | **468** | **145** | **60** | **50** | **27** | **22** | **13** |

Table 4: Number of feature matches with respect to image scale ratio obtained by MODS [27], KVLD [22] and our scale-invariant matcher on the *book-cover* image set.

As a versatile matching framework, our approach can be applied to any scale-invariant features. Matching results of ORB [35], SURF [3] and SIFT [23] features are visualized in Figure 6. Improvements upon all feature types are observed.

### 4.3. Ground-Aerial SfM Evaluation

In this section, we would like to evaluate the power of our large scale-invariant matching approach with SfM application. To this end, we collect a challenging Ground-Aerial dataset with large variations of viewpoints and view scales. As summarized in Table 5, the Ground-Aerial image set comprises 10467 high-resolution images, 4730 (4000×3000) for aerial and 5737 (4032×3024) for ground, and three main ground-aerial blocks. Following the matching pipeline in [1, 10], we start with image retrieval to select top 120 candidates for each query and then perform SIFT matching on candidate pairs. Our implementations of retrieval and feature matching are the same as Section 4.1 and 4.2. For comparison, the standard image matching framework involving the baseline retrieval technique [31] and SIFT matching [23, 46] in Section 4.1 and 4.2 is also conducted. Then, the canonical incremental SfM pipeline [41, 47, 36, 51] follows to recover cameras and structures. The matching performances are evaluated on the three individual blocks as well as the whole image set by counting

| Area | #Images | #Aerial images | #Ground images | Median scale ratio | Ours | | | Standard | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | #Ground-aerial retrieved pairs | #Ground-aerial matched pairs | Time[1] (ms) | #Ground-aerial retrieved pairs | #Ground-aerial matched pairs | Time[1] (ms) |
| Block A | 2309 | 822 | 1487 | 18 | 1339 | 919 | 5.24 | 104 | 0 | 78.42 |
| Block B | 3292 | 1945 | 1347 | 8 | 10946 | 7282 | 10.89 | 194 | 0 | 78.72 |
| Block C | 4866 | 1963 | 2903 | 10 | 10437 | 6129 | 10.95 | 646 | 0 | 94.81 |
| All | 10467 | 4730 | 5737 | N/A | 19717 | 13410 | 9.76 | 693 | 0 | 86.42 |

[1] Average matching time per pair.

Table 5: Statistics and comparative matching results on Ground-Aerial dataset. Matching pipeline is conducted on three individual blocks as well as the whole set. Our matching framework successfully produces a large amount of ground-aerial match pairs, while standard matching framework [31, 23] obtains none of these connections. And the match efficiency is greatly improved by our approach at the presence of large scale differences.

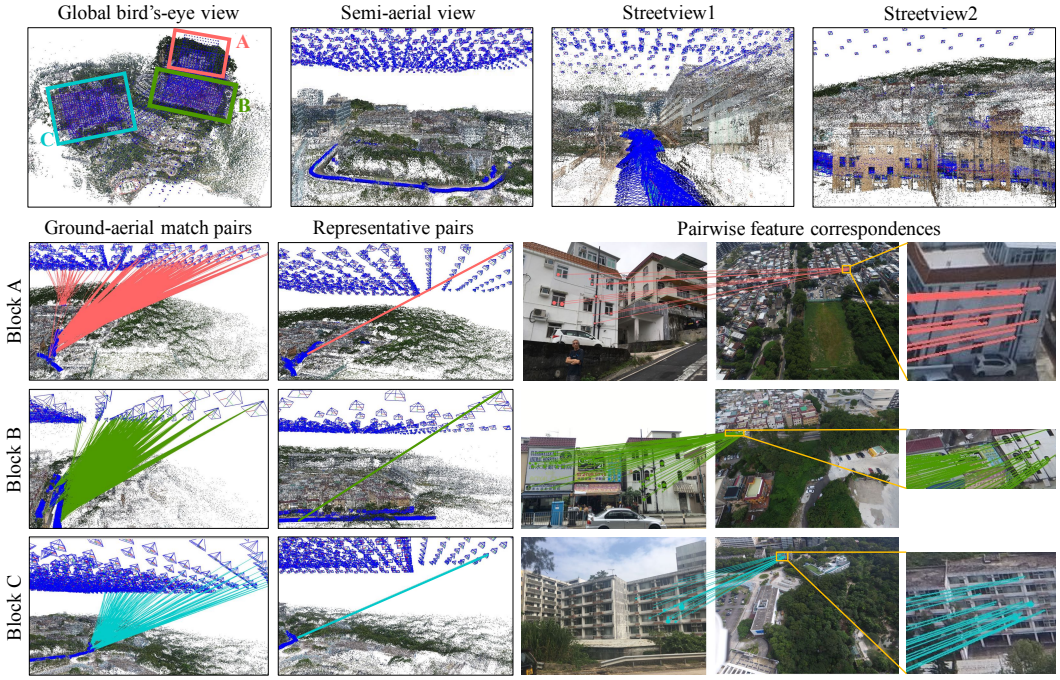


Figure 8: Visualization of ground-aerial matching results by our approach based on which complete structures and camera poses are successfully recovered. The first row shows the constructed camera poses and sparse points observed from different views. Below that, ground-aerial connections and sampled matching results are displayed in three colors for three blocks respectively.

the number of true positive ground-aerial image pairs that pass the retrieval stage and then the feature matching stage.

Our scale-invariant matching framework succeeds in attaining abundant matches between ground and aerial images with near decuple efficiency as shown in Table 5 and Figure 8. Thereby it ensures strong connectivity for complete and accurate fusion of ground and aerial views eventually. On the contrary, the standard image matching framework happens to retrieve a low quantity of weak ground-aerial image pairs. However, none of the pairs are successfully matched by SIFT matching [23, 46] subsequently, which leads to failure in giving complete SfM results.

## 5. Conclusion

In this paper, we have proposed a large scale-invariant image matching approach that manages to tackle the drastic scale differences robustly. It is composed of two progressive matching steps: first the scale level matching to find the related scale ranges following the scale consistency, and then the scale-aware matching to compute image similarity and find the exact feature correspondences. Its superior performance has been demonstrated via rigorous evaluations on retrieval and feature matching tasks as well as the very challenging ground-aerial fusion experiment. Improvement is expected if any more effective scale-space image encoding methods in Section 3.2 are available in future work.

**Acknowledgement** This work is supported by Hong Kong RGC 16208614, T22-603/15N, Hong Kong ITC PSKL12EG02, and China 973 program, 2012CB316300.

## References


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

[2] R. Arandjelovic and A. Zisserman. All about vlad. In *CVPR*, 2013.

[3] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *ECCV*, 2006.

[4] A. Bódis-Szomorú, H. Riemenschneider, and L. Van Gool. Efficient volumetric fusion of airborne and street-side data for urban reconstruction. *arXiv preprint arXiv:1609.01345*, 2016.

[5] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang. Spatial-bag-of-features. In *CVPR*, 2010.

[6] J. Cheng, C. Leng, J. Wu, H. Cui, H. Lu, et al. Fast and accurate image matching with cascade hashing for 3d reconstruction. In *CVPR*, 2014.

[7] O. Chum, A. Mikulik, M. Perdoch, and J. Matas. Total recall ii: Query expansion revisited. In *CVPR*, 2011.

[8] T. Fang and L. Quan. Resampling structure from motion. *ECCV*, pages 1–14, 2010.

[9] 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.

[10] 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*, 2010.

[11] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *PAMI*, 32(8):1362–1376, 2010.

[12] V. H. Hiep, R. Keriven, P. Labatut, and J.-P. Pons. Towards high-resolution large-scale multi-view stereo. In *CVPR*, 2009.

[13] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008.

[14] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *CVPR*, 2009.

[15] Y. Kalantidis, C. Mellina, and S. Osindero. Cross-dimensional weighting for aggregated deep convolutional features. In *ECCV*, 2016.

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[17] M. Lhuillier and L. Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *PAMI*, 27(3):418–433, 2005.

[18] S. Li, S. Y. Siu, T. Fang, and L. Quan. Efficient multi-view surface refinement with adaptive resolution control. In *ECCV*, 2016.

[19] X. Li, M. Larson, and A. Hanjalic. Pairwise geometric matching for large-scale object retrieval. In *CVPR*, 2015.

[20] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2):79–116, 1998.

[21] T. Lindeberg. *Scale-space theory in computer vision*, volume 256. Springer Science & Business Media, 2013.

[22] Z. Liu and R. Marlet. Virtual line descriptor and semi-local matching method for reliable feature correspondence. In *BMVC*, 2012.

[23] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[24] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004.

[25] A. Mikulik, O. Chum, and J. Matas. Image retrieval for online browsing in large image collections. In *SISAP*, 2013.

[26] A. Mikulík, F. Radenović, O. Chum, and J. Matas. Efficient image detail mining. In *ACCV*, 2014.

[27] D. Mishkin, J. Matas, and M. Perdoch. Mods: Fast and robust method for two-view matching. *Computer Vision and Image Understanding*, 141:81–93, 2015.

[28] J.-M. Morel and G. Yu. Asift: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2):438–469, 2009.

[29] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISSAPP*, 2009.

[30] M. Muja and D. G. Lowe. Scalable nearest neighbor algorithms for high dimensional data. *PAMI*, 36, 2014.

[31] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006.

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

[33] L. Quan. *Image-based modeling*. Springer Science & Business Media, 2010.

[34] F. Radenović, G. Tolias, and O. Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *ECCV*, 2016.

[35] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, 2011.

[36] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *CVPR*, 2016.

[37] J. L. Schonberger, F. Radenovic, O. Chum, and J.-M. Frahm. From single image query to detailed 3d reconstruction. In *CVPR*, 2015.

[38] Q. Shan, C. Wu, B. Curless, Y. Furukawa, C. Hernandez, and S. M. Seitz. Accurate geo-registration by ground-to-aerial image matching. In *3DV*, 2014.

[39] T. Shen, S. Zhu, T. Fang, R. Zhang, and L. Quan. Graph-based consistent matching for structure-from-motion. In *ECCV*, 2016.

[40] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[41] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring image collections in 3d. *SIGGRAPH*, 2006.

[42] C. Strecha, A. Bronstein, M. Bronstein, and P. Fua. Ldahash: Improved matching with smaller descriptors. *PAMI*, 34(1):66–78, 2012.

[43] C. Sweeney. Theia multiview geometry library: Tutorial & reference. `http://theia-sfm.org`.

[44] J. Wang, T. Fang, Q. Su, S. Zhu, J. Liu, S. Cai, C.-L. Tai, and L. Quan. Image-based building regularization using structural linear features. *TVCG*, 22(6):1760–1772, 2016.

[45] Z. Wang, L. Zhang, T. Fang, P. T. Mathiopoulos, X. Tong, H. Qu, Z. Xiao, F. Li, and D. Chen. A multiscale and hierarchical feature extraction method for terrestrial laser scanning point cloud classification. *ITGRS*, 53(5):2409–2425, 2015.

[46] C. Wu. SiftGPU: A GPU implementation of scale invariant feature transform (SIFT). `http://cs.unc.edu/~ccwu/siftgpu`, 2007.

[47] C. Wu et al. Visualsfm: A visual structure from motion system. 2011.

[48] J. Xiao, T. Fang, P. Zhao, M. Lhuillier, and L. Quan. Image-based street-side city modeling. *TOG*, 28(5):114, 2009.

[49] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[50] R. Zhang, S. Li, T. Fang, S. Zhu, and L. Quan. Joint camera clustering and surface segmentation for large-scale multi-view stereo. In *ICCV*, 2015.

[51] S. Zhu, T. Fang, J. Xiao, and L. Quan. Local readjustment for high-resolution 3d reconstruction. In *CVPR*, 2014.

[52] S. Zhu, T. Shen, L. Zhou, R. Zhang, J. Wang, T. Fang, and L. Quan. Parallel structure from motion from local increment to global averaging. In *ArXiv e-prints*, 2017.