ABSTRACT

Many well-known existing image matching methods are based on local texture analysis, and consequently have difficulty handling low-textured 3D objects, such as those man-made buildings and road networks in urban scenes.

In this paper, we propose our urban images matching method utilizing multiple novel clues. Specifically, we explore robust image features generated by interest regions and edge groups extracted from input urban images. Initial correspondences are established based on our features’ similarity measurement. Cost ratio, combined with global context information, is used to remove outliers. We believe our hybrid approach is more suitable than traditional texture features for such scenes.

The proposed method has been tested using real aerial urban photos under diverse viewing conditions. Experiments report that our registration success rate is nearly doubled compared with classical methods such as [2] and [3].

1. INTRODUCTION AND RELATED WORKS

Correspondence is one of the fundamental problems in image processing and computer vision. Notable progresses have been achieved in recent years concerning the image matching problem of high-textured planar objects. Many well-known existing methods are based on local texture analysis, describing patches around detected interest points.

The early work of Schmid and Mohr [1] proposed to use local greylevel invariants for orientation invariance, bringing public attentions to local features. It was demonstrated later (D.G. Lowe [2]) by the Scale Invariant Feature Transform (SIFT) method that rotation and scale invariance can be achieved by first using difference-of-Gaussian function to detect stable interest points, then constructing local region descriptors using histograms with assigned orientations and detected scales. Different from the fixed-shape interest regions in SIFT, [3] presented two methods extracting affine invariant regions. The later work of Ke and Sukthankar [4] applied Principal Components Analysis to image gradient patches replacing the smoothed weighted histograms in SIFT.


Above methods, although utilizing various feature detectors and descriptors with different performance strength, are essentially built upon the same kind of feature: textures of local image patches, which have been intensively verified for planar objects with high-textured surfaces. However, for urban scenes of 3D buildings without rich textures, experiments indicated their performance is no longer reliable due to various reasons such as the lack of appropriate textures and the violation of general affine transformation in such scenes, or the difficulty in modeling textures for non-planar objects under complex 3D viewpoint changes, especially for boundary regions.

In this paper, we propose to combine two different kinds of features: the interest regions and the edge groups for the challenging urban image matching problem. The regions we are interested in (ROI) ideally are representing conceptually meaningful parts of buildings that are well-distinguished from neighborhood areas. Because regions and edges are closely related in nature, instead of treating the two feature tracks independently, we construct them in an interleaved manner. The initial edge detection results serve as references for ROI extraction component, while the extracted ROI will guide the meaningful grouping of edges.

Experimental results for several major U.S. cities’ datasets including Atlanta, Baltimore, Denver and Los Angeles, comparing with other state of the art image matching methods, have demonstrated the effectiveness of our proposed approach.

2. FEATURE DETECTION AND DESCRIPTION

This section presents detection and description details of the two kinds of features we combined: ROI and edge groups.

2.1. Edge Detection and ROI Extraction

Our feature detection component for the ROI track is a region-growing process utilizing two assistant information...
maps (AIM) and learning techniques.

First of all, two AIM are constructed: vege-map and edge-map. The input images are also selectively smoothed during the constructions. For vege-map, we identify pixels dominated by green channel and possibly vegetation areas.

![Image](image_url)

**Fig. 1.** Effectiveness of in-region edges and selectively smoothing: (a): the initially detected edges (left) and one patch (right) showing true edges (bright edges) and in-region edges (grey edges); (b): the same area before (left) and after (right) selectively smoothing.

There are two kinds of edges in our edge-map, the true edges and the in-region edges. Among the initial edges returned by Canny operator, many are not the actual boundaries of ROI we are interested in (true edges) but rather edge responses within those regions (in-region edges). Since our whole registration process is a combination of region-driven and edge-driven, it is meaningful to distinguish those two kinds of edges from the very beginning. One direct outcome of in-region edges is selectively smoothing. The color and intensity of each confirmed in-region edge pixel will be replaced by the average of its non-edge neighbors. The selectively smoothing helps us eliminate those in-region details which otherwise will compromise ROI extraction performance.

During the region growing process, there are three expansion/boundary conditions for the current pixel (p_c):

First, p_c must not belong to the vege-map. And if it belongs to an in-region edge, it will still pass this test though with a penalty to its confidence measurement. If it belongs to a true edge, it will be neither accepted nor further expanded.

Second, we use the fractal geometry [8] to constrain the ROI boundaries. In our case, we use intensity I(p) as measurement and the linearized equation of fractional Brownian motion model for neighboring pixels (p_r) is:

\[
\ln(E[I(p_r) - I(p_c)]) = \ln(k) + \ln(|p_r - p_c|)
\]

Least-squares linear regression is used to compute the optimized \(\bar{H}\) and \(\bar{k}\). The overall fractal error (OFE) for p_c is computed as the root mean square of the difference between the actual and estimated fractal values from all its neighboring pixels. Only pixels with sufficient high OFE will pass this expansion requirement.

Last is the dynamic intensity range defined by the upper and lower bounds, expanded simultaneously with the region growing process. We use a tolerate threshold as the expansion limit, softened based on the current area to handle the case when the current point falls into a small distinct region contained in a large region we are interested in.

### 2.2. Edge Grouping

First, membership labels are assigned to each edge pixel based on its spatial relationship with its neighboring ROI. When multiple ROI are sufficiently close to one edge pixel (e.g. the distances to the closest and the second closest ROI are similar), multiple labels are assigned.

![Diagram](diagram_url)

**Fig. 2.** Overview of the edge grouping component

Second, separated edge pixels are grouped into short straight lines (called “edgelets”). Even when a long line could be detected, our algorithm will still return a number of short lines for it because edgelets are more flexible for describing complex boundaries, especially curved ones. The other reason is the convenience of forming dense descriptors starting from edgelets. Traditional sparse descriptors for long lines or curves, simply based on length, angles and curvatures etc., are insufficiently robust for challenging urban scenes because they frequently encounter broken or missing parts from one view to the other, due to shadows, occlusions and viewpoint changes, while our experiments demonstrated that dense descriptors from edgelets are considerably more tolerated to those factors.

A voting histogram is constructed for each edgelet recording the number of votes each label receives. The label of the whole edgelet is set to be the label of the histogram’s first (highest) peak. Edgelets with the same label will be grouped together. A second label could be assigned if the height of the second peak is relatively close to the first.

Finally, each edge group is stored as one or multiple ordered edgelet lists (OEL). One ordered list initially contains only one edgelet in the group with an end point closest to its bounding box. Other edgelets with an end point sufficiently close to the outgoing end point of the current list are iteratively added to the list, until the list returns to the initial edgelet or no close one could be located any more. Then all the edgelets in the list are removed from the group and the process repeats until the group is exhausted.
2.3. Description of Edge Groups and ROI

Edge groups are described based on their OEL using a set of weighted 2D histograms. Our concerns for descriptors include: tolerance to location errors from edge detection and grouping inaccuracy, geometric distortions commonly seen in urban photos, and the ability to provide dense descriptors for point-to-point correspondences. Based on those considerations, the descriptors we use are similar to Shape Context [9] in global structure but some modifications and improvements are introduced.

First of all, the basic description primitives in our case are edgelets instead of contour points in [9]. Because our edgelets are already organized in a spatial continuous order, we can uniformly sample the whole OEL to produce a fixed number of virtual anchor points (NVAP=30–50 in our experiments) serving as description centers. They are described by the relative logarithm distances and angles from all edgelet midpoints (index 0<k<Nedgelet) in the current list (index i), weighted by the edgelet lengths, thus producing a 2D weighted histogram for each VAPj.

Second, rotation invariance is achieved by using the edgelet directions for midpoints and computing tangent vectors for each VAP, so that a rotation relative frame can be used instead of the absolute one. Next, scale invariance is enhanced by normalizing the distances using the size of the current OEL bounding box, which is shown by our experiments to be more stable than other criteria under edge grouping errors. Finally, to improve the performance in terms of over-grouping of edges, which occasionally happens at heavily urbanized areas, we construct additional partial descriptors using continuous edgelets subsets containing larger number of corners.

The ROI description is similar with two major differences. Instead of VAP, actual contour points are used. Moreover, what the histograms contain is no longer edgelets, but sampled contour points excluding the current description center. As a result, the histograms are no longer weighted.

3. ESTABLISHING CORRESPONDENCES

The proposed system establishes initial matchings for the edge group track and the ROI track respectively, after which the two sets of initial matchings are merged together and outliers are removed to produce final correspondences.

In order to establish correspondences for edge groups, for the NVAP histogram descriptors (H1) of one OEL in one input image, the descriptors (H2) of every OEL (index i) in the other image are sequentially scanned and the most similar one are selected as matching. We measure the similarity of two OEL as the minimum average histogram distance (“matching cost”) of their corresponding VAP:

$$\min_{0 \leq k < N_{edgelet}} \frac{1}{N_{VAP}} \sum_{i=0}^{N_{VAP}} \sum_{j=0}^{N_{edgelet}} H_1(j) - H_2((j+k) \mod N_{VAP})$$

Each OEL matching can provide NVAP point-to-point correspondences. Similarly, the ROI track can also produce a large number of initial correspondences based on ROI matching results.

We define "cost ratio" for each matching as the matching cost ratio of its best matching over the second best. Low cost ratio indicates a more distinctive matching and high matching confidence. Matchings with high cost ratio are discarded. Next, we apply RANSAC to locate the best consistent subset among the remaining matchings. Finally, a global transformation (H_T) is estimated from the selected consistent subset. The correspondence for any image point can be roughly estimated using H_T.

4. EXPERIMENTAL RESULTS

The proposed method has been tested using both real and synthesized challenging aerial photos of several major cities’ datasets. Around 30% of our testing images have viewing directions almost perpendicular to the ground, the rest are oblique views with various unknown 3D camera rotations and zoom-levels.

4.1. Invariance

First, robustness to scale changes comes from our distance normalization and by placing edgelet lists or ROI of different scales into histograms with a fixed number of r bins. Second, rotation invariance is achieved by constructing descriptors in relative frames that automatically turns based on directions of edgelet and tangent vectors. Third, global lighting changes generally have no influence on our method which doesn't use any texture-based features. Strong local lighting changes such as long shadows in heavily urbanized areas could notably influence region and edge detections. Certain degree of robustness can be achieved through partial matching. Next, geometric distortions are tolerated by histogram-based dense
descriptors. Finally, it naturally handles occlusion because although we use high-level features, each feature is only related to a local area of the scene, while global constrain is applied only at the end.

Experimental results indicate ROI track generally provides better correspondences for areas more homogenous in nature, while edge group track is more suitable for parts lie between those areas. Through merging initial matchings from both tracks and refine them together using global context information, our method is able to adapt between those two kinds of features automatically.

4.2. Registration Accuracy and Success Rate

Our method is compared with existing approaches including [2], [4], [6] and [7]. We found the original SIFT, although comparatively slow to compute, produced best results in terms of both accuracy and success rate among the existing methods we tried.

During each experiment, the distance ratio [2] of SIFT varies from 0.6 to 0.9 and the best result is kept. We say one registration is a successful one if the refined inliers can recover a transformation roughly aligning the two input images properly (through visual inspection of corresponding buildings). The accuracy is measured by average pixel errors of point-to-point correspondences after RANSAC compared with manually labeled ground truth. Primarily due to the basic lack of appropriate textures in urbane scenes, throughout our experiments, the registration success rate of SIFT is below 40%. As comparison, our method fusing two different kinds of features achieves a success rate of 86%. Approximately 23% of the total tests are successfully registered by both methods. Most correct matchings returned by SIFT are on high-textured planar grounds, while our method is able to register 3D man-made structures using region shapes and edge groups.

Concerning the registration accuracy, as a local texture based approach, SIFT has sub-pixel level accuracy. Our experiments report that for those image pairs SIFT can successfully register, the average pixel error is within 2 pixels for matchings after RANSAC. The same error of our method is 5-10 pixels. The reason primarily comes from the difficulty of locating exact pixel locations inside high-level features we use, intensified by issues such as shadows, imperfect segmentation and line grouping, etc. The errors for final dense matchings propagated using $H_T$ could be as large as half a building’s size due to the fast but less accurate global transformation we are currently using. The current results are accurate enough for building recognition or rough alignment purposes. For applications demanding higher accuracy, refinement processes could be added.

5. REFERENCES