Feature Fusion for Registering Wide-baseline Urban Images

Quan Wang, Suya You

Computer Science Department
University of Southern California
Los Angeles, California, U.S.A.
{quanwang, suyay} @graphics.usc.edu

Abstract

This paper proposes a novel automatic image registration method designed for urban scenes containing tall 3D objects under significant viewpoint changes. Our method is based on the fusion of two kinds of features: the interest regions and the edge groups, which we believe are more suitable than conventional local textures for such scenes. Many existing methods face matching difficulty either because of the lack of appropriate textures in urban scenes or due to the difficulty in modeling textures for non-planar objects under complex 3D viewpoint changes, especially for boundary regions. Throughout our registration process, the two "feature tracks" interactively support and constrain each other, producing a large number of initial point-to-point correspondences, which are later refined utilizing global context information.

The proposed approach has been tested using urban aerial photos of diverse viewing conditions. Experimental results for challenging inputs have demonstrated its effectiveness.

1. Introduction

Image matching is a fundamental task in computer vision, used to correspond two or more images taken, for example, at different times, from different aspects, or different sensors. Image matching is also a core of many image/video processing systems as well as newly-emerged exciting applications, such as image-based content retrieval (CBIR), mobile augmented reality (MAR), 3D scene modeling and texturing, and vision-based personal navigation and localization, etc.

Concerning planar objects with high-textured surfaces, notable progresses have been achieved in recent years. Many well-known existing methods (e.g. [1], [2], [3] and [4]) are based on local texture analysis around detected interest points. However, the classical problem is particularly challenging for urban areas. When it comes to 3D non-planar low-textured objects, such as buildings in typical urban scenes, traditional texture-based approaches tend to face great difficulty either basically because of the lack of appropriate textures (e.g. man-made structures commonly have limited textures or contain textures in repeated patterns which easily confuse texture-based methods), or due to the difficulty in modeling textures for non-planar objects under complex 3D viewpoint changes, especially for boundary regions.

Despite the technical difficulties, recently there has been a growing need for registering urban images under diverse viewing conditions. For example, with the recent development of remote sensing technology [13] and high-resolution stereo equipments [14], high quality depth data for urban areas can be efficiently obtained, resulting un-textured city models. Although partial textures could be obtained through nadir images with georeference provided by positioning hardware, the problem of registering oblique or even ground images still remains largely unsolved.

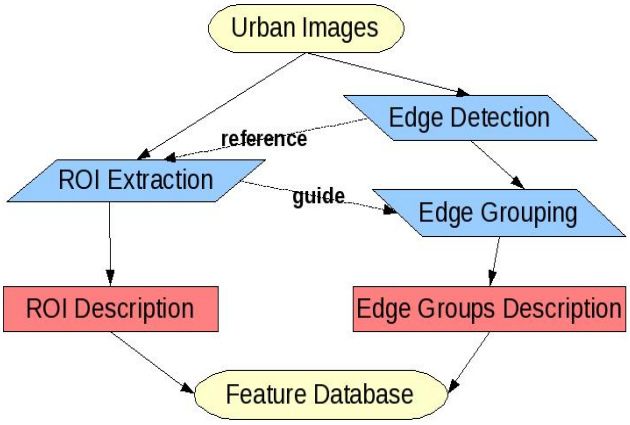


Fig. 1. Major components for feature detection and description containing ROI track (left) and edge groups track (right)

In this paper, we propose our automatic urban image registration method utilizing two kinds of features: the interest regions and the edge groups. We believe that a hybrid approach to fuse these two features is more suitable than texture-based features for urban scenes. The regions we are interested in (ROI) represent conceptually meaningful parts (e.g. roofs) of buildings that are well-distinguished from neighborhood areas. Because regions

and edges are closely related in nature (e.g. many edges are boundaries of regions and ROI typically have less internal edges), instead of treating the two feature tracks independently, we construct them in an interleaved manner (overview in figure 1). The initial edge detection results serve as references for ROI extraction component, while the extracted ROI will guide the meaningful grouping of edges. Dense and robust descriptors for ROI and edge groups are constructed respectively. Initial correspondences are established based on our features' similarity measurement. Finally, cost ratio, combined with global context information, is used to remove outliers.

This paper primarily focuses on how the two kinds of high-level features are interactively fused together in order to provide successful registration in challenging urban areas where many other state-of-the-art methods fail. Low-level details such as the detection and description of individual feature are only briefly mentioned in this paper or skipped.

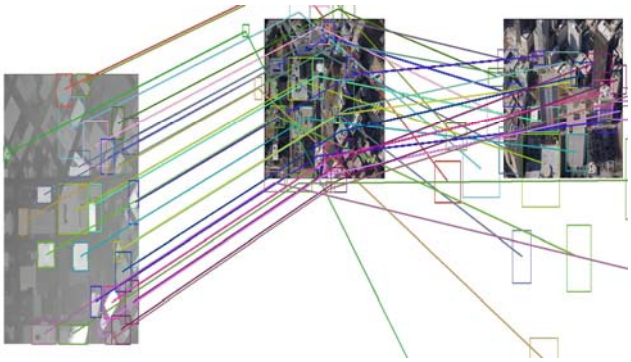


Fig. 2. One result from the 2D-3D registration system visualized by the bounding boxes and centers of all ROI' point-to-point correspondences. The proposed method is used to register nadir (middle) to oblique images (right).

The proposed method has been tested using aerial urban photos under diverse viewing conditions. Furthermore, our method has also been successfully integrated into a 2D-3D registration system for Light Detection and Ranging (LiDAR) data (figure 2).

2. Related works

The early work of Schmid and Mohr [1] introduced a rotationally invariant descriptor for local image patch, which is based on local greylevel invariants. Local grayscale invariants are automatically computed around interest points.

The ground-breaking work of D.G. Lowe [2] demonstrated that rotation as well as scale invariance can be achieved by first using difference-of-Gaussian function to detect stable interest points, then construct the local region descriptor using assigned orientation and several histograms. Efficient searching technique [5] was also proposed for the generated high-dimensional feature vectors.

The proposed SIFT method produced significant influence on later works. For example, different from the fixed-shape interest regions in SIFT, [3] presented two methods extracting affine invariant regions from a single image covering identical parts of a scene. Ke and Sukthankar [4] applied Principal Components Analysis (PCA) to an image gradient patch in order to reduce the descriptor's dimensionality instead of using the smoothed weighted histograms like in the original SIFT. Other variations of SIFT method include: Bay et al. proposed a fast implementation of SIFT using integral images [18]. The proposed Speeded Up Robust Feature (SURF) is built upon other experiments-verified and successful detectors and descriptors, especially SIFT, but simplify those steps to the essential. The new descriptors can be constructed and matched more efficiently.

Another line of research is to combine the information of multiple views and train the system in an offline stage so that it will learn the main characters of the target object under different viewing conditions. Consequently the online matching process can achieve interactive [8] or even real-time [9, 10] speed. Instead of the expensive histogram-based description [2] or pixel extraction from irregular regions [3], image matching is treated as a classification problem by matching inputs to compact representations of training views using learning [9] or projection [10] method.

Although robust to various geometric distortions, when it comes 3D non-planar objects with large viewing direction changes, typical scenarios found in modern urban photos, all the above mentioned image features based on local texture analysis such as SIFT descriptors and its many variations have very limited power due to various reasons such as the diversity of texture appearances near object boundaries when the general affine assumption is violated.

To tackle the problem, targeting on urban scene modeling and rendering applications, many recently developed methods first reconstruct sparse or dense 3D point clouds from image sequence or video clip, and then use high level features (e.g. 3D edges, intersection of perpendicular 3D lines) to establish correspondences.

Zhao, et al. [11] use motion stereo to recover dense 3D point clouds from continuous oblique video clip. ICP (Iterative Closest Point) algorithm is used to register those recovered 3D points with initial alignments provided by positioning hardware such as GPS (Global Positioning System) and IMU (Inertial Measurement Unit).

Liu, et al. [19] apply structure from motion technique to a sequence of images to produce sparse point clouds of an urban building. A novel approach for automatically aligning dense and sparse point clouds is proposed to register images from various viewing directions onto the same 3D model.

Both [11] and [19] try to use multiview geometry methods to recover 3D point clouds from image sequence or video clip. The first limitation is that such multiple views of the interest object/scene might not always be readily available. Second, as the first step of 3D reconstruction methods using multiview geometry, correspondences among 2D images need to be established, which is a very challenging Computer Vision problem by its own especially for wide base-line urban scenes. Last but not the least, even suppose a number of perfect correspondences can be obtained, traditional stereo or structure from motion techniques still tend to produce inconsistent and noisy results.

The later work of Ding, Lyngbaek and Zakhor [12] assumes clean 3D building models of urban areas as inputs. They detect 2D orthogonal corners (2DOC) and use them as primitives to match urban aerial images to 3D city models. Similar to [11], it is also very important that their matching process is properly initialized (average initial error smaller than 1.25% of the input image size, according to our own visualization of similar datasets) utilizing information provided by positioning hardware.

However, we can not assume such hardware-provided highly accurate initial estimation is available for general wide-baseline urban registration problem, e.g. for historic pictures or photos from common users. Without proper initialization, registration methods based on ICP or local search of orthogonal corners could not be sufficient.

3. Edges and regions detection

This section presents how our matching primitives: edges and interest regions are detected and extracted from input urban photos respectively.

For the ROI extraction component, our ultimate goal is to extract, from urban photography, building's dominant and most external contours. Towards this goal, our aerial image segmentation algorithm is built on region-growing process utilizing assistant information maps (AIM) and learning techniques. Edges are also detected and extracted during the same process, which are going to be grouped and described later in our registration system. Considering the many challenges of aerial image segmentation problem and the particular need of our application area, we combine the traditional edge-driven and region-driven segmentation approaches together and introduce several techniques in order to extract the interest regions our registration system needs.

First of all, we identify edge pixels using Canny operator. Among the initial edges returned by Canny operator, many of them are not the actual boundaries of interest regions (true edges) but rather edge responses within those regions (in-region edges) due to slopes or textures of roofs, items like air conditionings on some building's top, or even noises from image sensors. The existence of those in-region edges is one primary reason why many segmentation algorithms tend to over-segment.

For urban scenes with regular buildings, an edge pixel is deemed as a part of a true edge unless its neighboring horizontal or vertical non-edge pixels have similar hues. Here we choose to use HSV instead of RGB color space because neighboring pixels of either true or in-region edges tend to be affected by different lighting and shadow conditions, and HSV color space is known to be more robust under such circumstance.

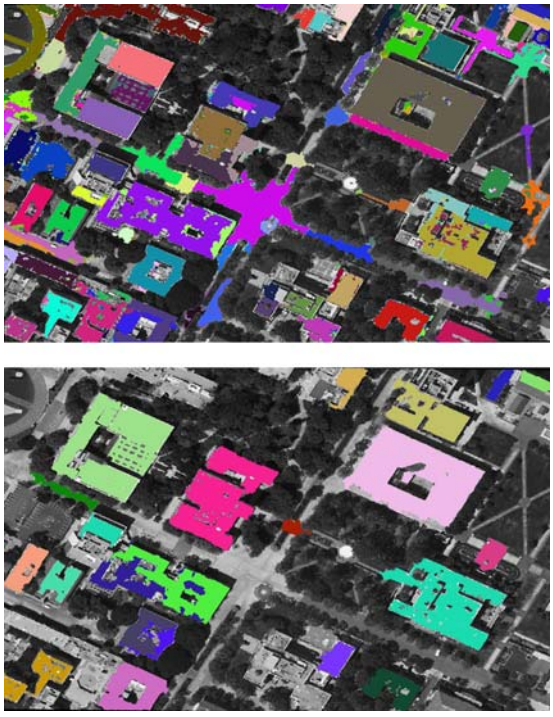


Fig. 3. Top: color-coded initial segments of one input urban image. Bottom: the final segmentation results for the same image after region regulation and merging.

Second, using the above classification of edges as references, pixels of the input image are selectively smoothed. Areas around true edges are blurred while areas around true edges are preserved. This stage tries to remove the noises and unnecessary details within interest regions while preserving the actual boundaries of those regions at the same time.

Next, the actual region-growing process has a similar structure as the watershed algorithm [20] but is

additionally constrained by our newly proposed parameters including the color preferences and two automatically-adjusted dynamic intensity thresholds, which are determined by a combination of multiple cues such as the fractal error metric [15] and the assistant information maps. During this region-growing process, each initially segmented region is also associated with an uncertainty attribute.

Fourth, the initial segmentations produced by the above region-growing process (one example in figure 3) will be pruned by our histogram-based region regulation, which is to prune those regions that are too irregular to become a building region or a part of such region. For similar tasks, many previous works define different prototypes of common building shapes and try to fit one of those prototypes to each segmented regions. Such methods will fail when encountering a new shape that is not in the pre-defined library, for example, a unique shaped building or parts of a regularly-shaped building produced by over-segmentation. In contrast to those model-fitting approaches, our histogram-based region regulation using machine learning techniques is more efficient and generalized because no pre-defined models are needed. During the regulation process, a confidence attribute is computed for each ROI.

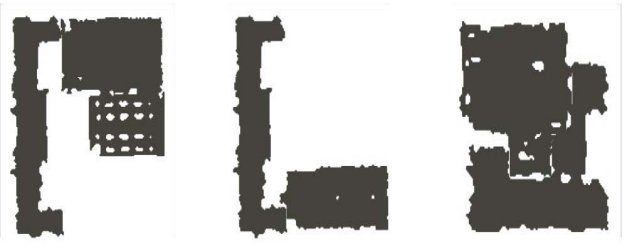


Fig. 4. additional interest regions as the combination of spatially close initial regions

Finally, due to distinct appearances of some building's different parts or overlapping shadows, the same building might be segmented into a few or even several regions, thus compromising the performance of the final region matching component. In order to further handle over-segmentation problems, some of the remaining interest regions will be iteratively merged together to form additional regions according to their spatial relationship, uncertainty and confidence attributes. It should be noted that no region is removed or overwritten during each iteration. We believe that all the original regions should be kept and merged regions (new regions) should be saved as additional regions due to the fact that segmentation of urban image is such a complicated task that one decision good for one scene may bring unwanted results for another. Therefore, it is safe to keep all the original and merged regions even some of them might be the fragments of correct segmentation.

Figure 4 shows three examples of merged new additional regions after the first iteration. They are appended to the region list as additional interest regions.

4. Edges grouping

One crucial step before feature description is to organize initially detected edges into conceptually meaningful groups, for example, representing boundaries of roofs or sides of buildings or a parking lot, etc.

Overcoming the disadvantages of traditional primitive-based edge-grouping approaches, our method uses the previously extracted ROI to guide the grouping of edges, just like initially detected edges are used by the ROI extraction component to constrain the region-growing process.

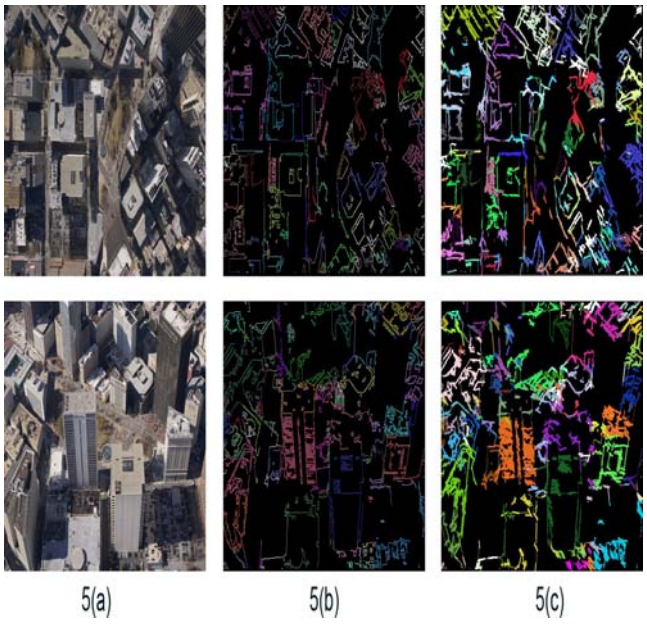


Fig. 5. (a): input urban images; (b): labeling of edge pixels; (c): labeling of edgelets

First, one or multiple group membership labels are assigned to every edge pixel according to its spatial relationship with its neighboring ROI. After the initial membership assignment, all edge pixels are grouped into short straight lines (“edgelets”). Edgelets are used as matching primitives in our edge matching track because they are more flexible for describing complex shapes. Furthermore, based on our edge feature description method, edgelets naturally produce dense descriptors, resulting more robust matching performance.

Next, we construct a voting histogram for each edgelet based on the membership voting of each edge pixel on it. The highest peak of the histogram will provide the final membership label for that edgelet or when other peaks sufficiently close to the highest one exist, multiple final labels will be assigned. Figure 5 shows the edge grouping results with different colors representing different groups. Notice only one color (group membership) could be visualized for each edge pixel or edgelet.

Finally, each edgelet group is stored as one or multiple ordered edgelet lists (OEL), representing connected boundaries of the group.

5. Feature description and matching

We describe each ROI using 2D histograms of the relative distances and angles of its sampled contour points, similar to the shape context [17]. To achieve orientation invariance, they are computed in relative frame, which is constructed by computing tangent vectors for each pixel [6]. Other adaptations we introduced to the shape context include: additional partial descriptors are formed using continuous subset of the original contour points containing a larger number of corner points. Scale invariance is enhanced by distance normalization using ROI bounding boxes' sizes, which we believe is more stable under imperfect segmentations.

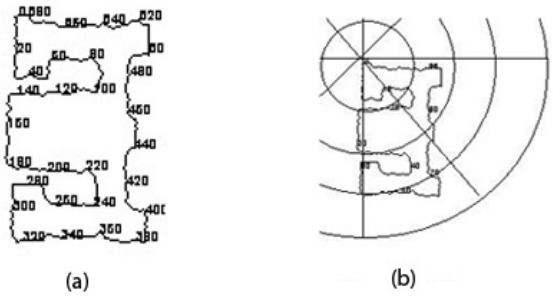


Fig. 6. (a): initial labeling of the ROI contour points; (b): ROI with sampled contour points placed into log-polar system.

The feature description for OEL is similar to ROI. The major difference is that the locations of contour points are replaced by the mid-points of edgelet, which are also weighted by the edgelet lengths.

To establish correspondences, take the ROI track as an example, we measure the similarity of two ROI as the minimum average histogram distance (matching cost) of their corresponding sampled contour points. Because in our case, those points represent the most-external contours of ROI in relative frame, instead of solving the expensive general bipartite matching problem, searching complexity becomes a low constant by first ordering the contour points in counter-clock manner and only searching close locations for correspondences. Once one point's matching is determined, the rest points are automatically corresponded. The same matching process is also applied to the OEL track. After our system obtains image correspondences from the edge-groups track and the ROI track respectively, the two sets of initial matchings are merged together and outliers are removed to produce the final point-to-point correspondences.

We introduce "cost ratio" as the ratio of best matching's cost over the second best matching's. A lower cost ratio indicates a ROI correspondence with higher confidence. As is shown in figure 7, rectangle ROI is generally ambiguous and produces higher cost ratio because many ROI in urban scenes have similar shapes, while ROI of special shapes will produce lower cost ratio, therefore higher matching confidence.

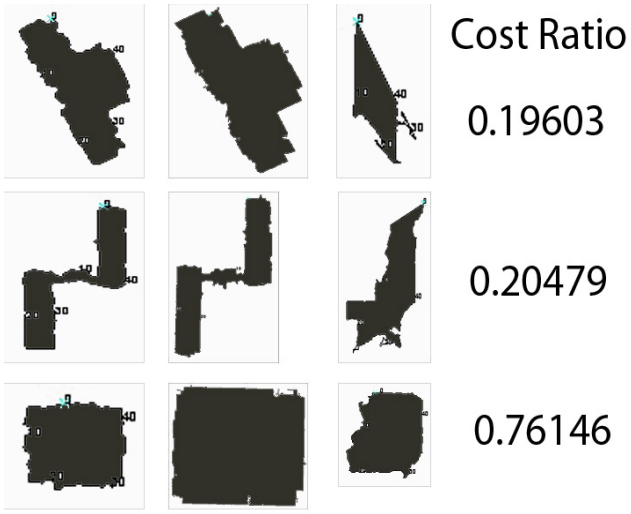


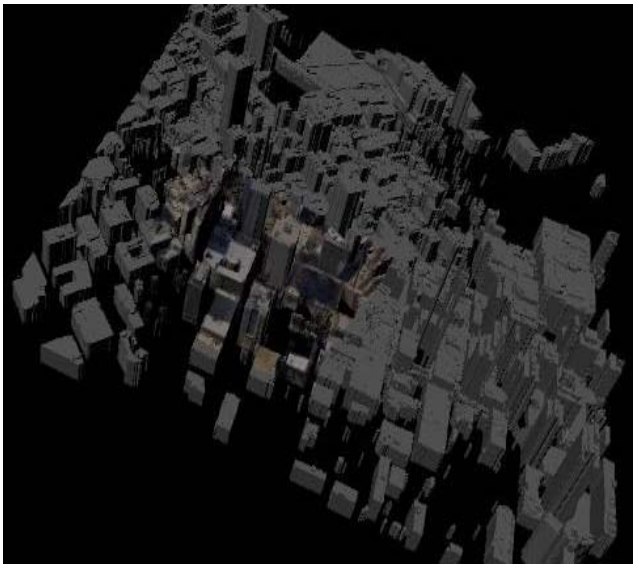
Fig. 7. The 1st column is ROI contours extracted from one urban image. The 2nd and 3rd column are the best and second best matching from another image of roughly the same area.

With the help of cost ratio, we apply an RANSAC-like method to remove outliers among the initial ROI correspondences. The consistent set is used to compute a global transformation (T) and with it the matchings are propagated across the entire scene. It is during this process, global context is implicitly taken into consideration.

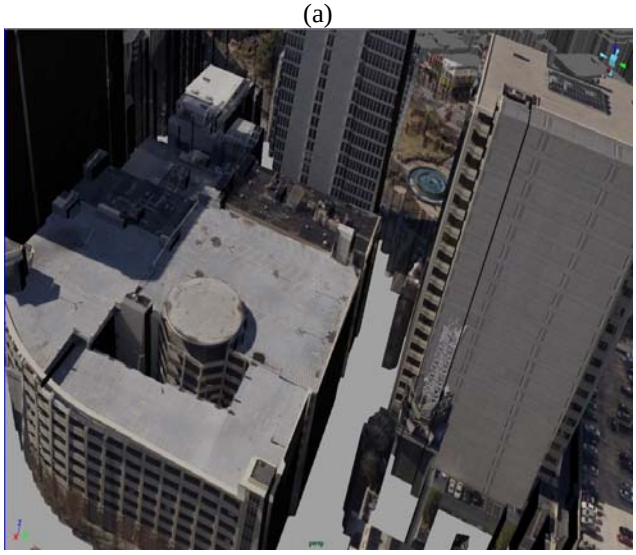
6. Experimental results

The proposed urban image registration method has been tested using both real and synthesized aerial photos of several major cities' urban areas with diverse viewing conditions. Around 30% of our testing images have

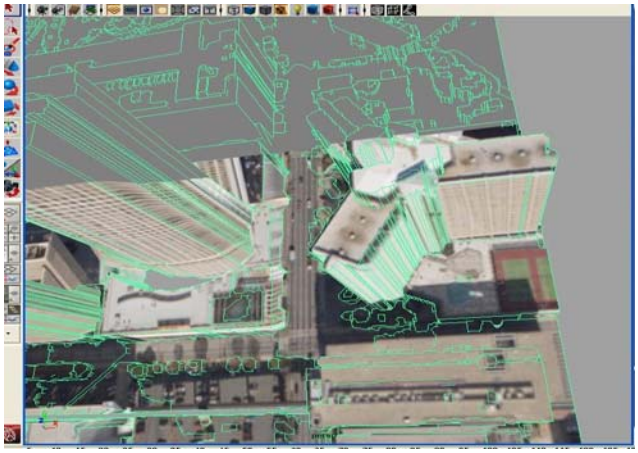
viewing directions almost perpendicular to the ground, the rest are oblique views with various and unknown 3D camera rotations and zoom-levels. Most of our tests involve one top-view and one oblique-view images. Besides distortions from real viewpoint changes, we also apply geometric transformations to generate some challenging synthesized testing images.



(a)



(b)



(c)

Fig. 8. The proposed method applied to urban modeling and rendering tasks.

Furthermore, we have applied the proposed techniques to a registration system for photorealistic modeling and rendering of urban scenes. First, wide-baseline urban images are registered with nadir images using the feature fusion technique proposed in this paper. Next, nadir images are registered with 3D city models using the method proposed in [21]. Consequently the input urban images become “indirectly” registered with 3D models. Figure 8 shows some representative results of the system without utilizing positioning hardware throughout the whole registration process.

6.1. ROI Extraction Results

ROI extraction is the foundation of the ROI matching track and also guides the edge grouping component. Our method achieves better results for areas with well-separated buildings (e.g. Los Angeles images in 9.a) than heavily urbanized areas (e.g. the Atlanta image in 9.b). Overall, a minimum of 30%-50% of all interest regions can be correctly extracted. The problems of segmentation breaking and leaking can be alleviated through partial descriptors. Nonetheless, it is neither reasonable today nor necessary to request perfect image segmentation. We believe that the important thing is how to make the best use of imperfect segmentation results.

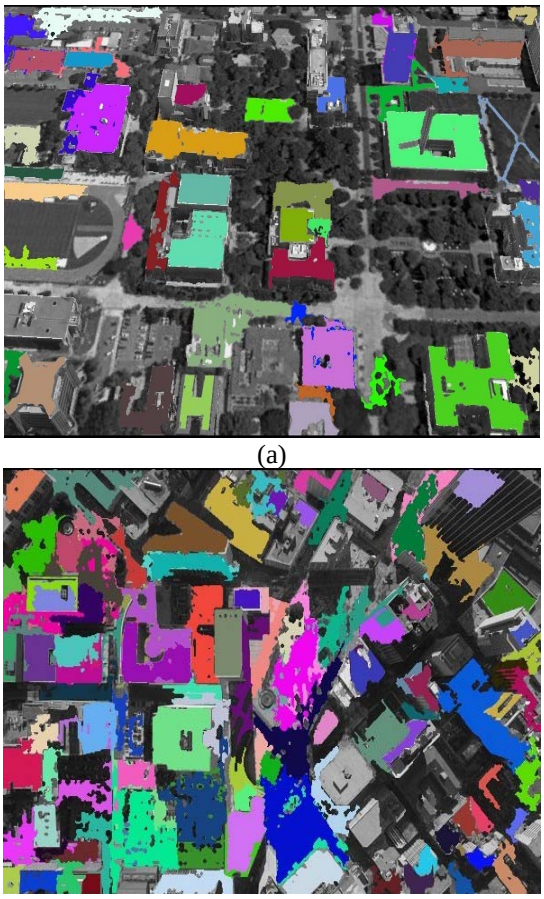


Fig. 9. Color coded ROI extraction results

6.2. Invariance

Our proposed feature fusion method achieves scale invariance through scale-invariant feature description and distance normalization. Rotation invariance comes from relative frames that automatically turns based on directions of edgelet and tangent vectors. Geometric distortions are tolerated by histogram-based dense descriptors. Last, as a local feature based approach, it naturally handles occlusion in large urban scenes.

Experimental results using various cities' datasets of difficult scenarios with significant viewpoint change have demonstrated the effectiveness of the proposed system in terms of common challenging factors of urban scenes. Additional, our results also indicate that the proposed approach can tolerate large perspective distortion and is able to handle noisy and extremely low resolution inputs.

6.3. Registration Accuracy and Success Rate

Our registration method using feature fusion is compared with existing approaches including [2], [4], [9], [10] and [18]. 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.

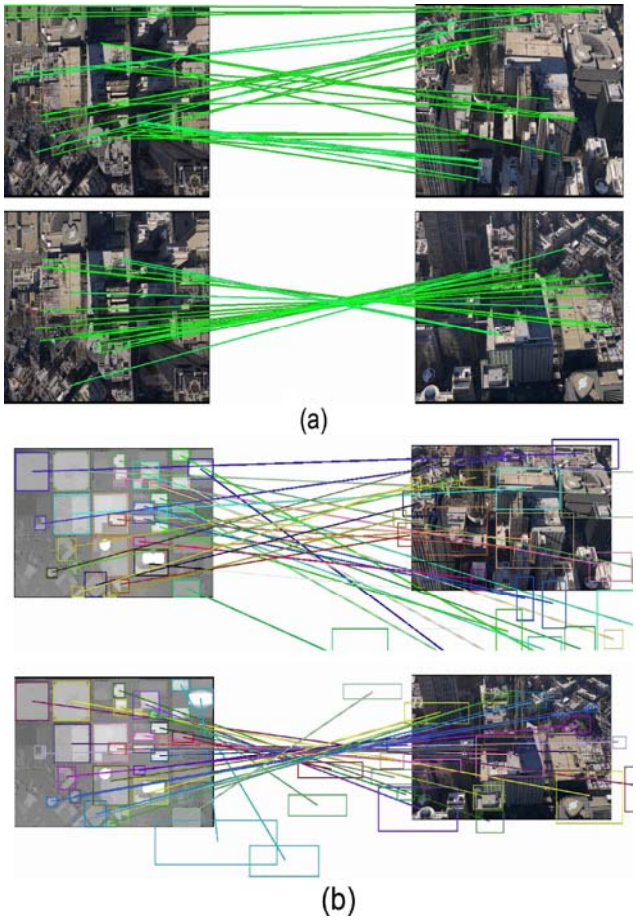


Fig. 10. Correspondences after RANSAC (a), which are used to generate (b) for 2D-3D registration application (see fig. 2), roughly recovering the scene poses and building locations

Primarily due to the basic lack of appropriate textures in urbane scenes, throughout our experiments of four cities' datasets, the registration success rate of SIFT is below 40%. As comparison, our method fusing two kinds of features achieves a success rate of 86%.

Concerning the registration accuracy, as a pixel and local texture based approach SIFT has sub-pixel level accuracy. Experimental results 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 current results are accurate enough for building recognition or rough alignment purposes. For

applications demanding higher level of accuracy, refinement processes such as those in [11] and [12] could be added.

7. Conclusion

This paper presents our novel image registration method for urban scenes based on fusion of interest region features and edge group features. Intensive experimental results demonstrated the potential of our approach. The proposed method has also been successfully applied to urban scene rendering tasks.

References

[1] C. Schmid and R. Mohr, Local grayvalue invariants for image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 530–534

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

[3] T. Tuytelaars and L. V. Gool, Matching widely separated views based on affine invariant regions . International Journal of Computer Vision, 59(1):61-85, 2004

[4] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. IEEE Conference on Computer Vision and Pattern Recognition, 2004

[5] J. S. Beis and D. Lowe, Shape indexing using approximate nearest-neighbor search in high-dimensional spaces, IEEE Conference on Computer Vision and Pattern Recognition, 1997

[6] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. Technical Report UCB//CSD-00-1128, UC Berkeley, January 2001

[7] Q. Wang and S. You, Feature Selection for Real-time Image Matching Systems, Proceedings of 19th International Conference on Pattern Recognition, Tampa, Florida, 2008

[8] V. Lepetit, J. Pilet and P. Fua, Point matching as a classification problem for fast and robust object pose estimation, Computer Vision and Pattern Recognition, 2004

[9] V. Lepetit and P. Fua, Keypoint Recognition using Randomized Trees, Transactions on Pattern Analysis and Machine Intelligence, Vol. 28, Nr. 9, 2006, 1465 - 1479

[10] Q. Wang, S. You, Real-Time Image Matching Based on Multiple View Kernel Projection, IEEE Conference on Computer Vision and Pattern Recognition, 2007

[11] W. Zhao, D. Nister, and S. Hsu. Alignment of continuous video onto 3d point clouds. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8):1305 - 1318, 2005

[12] M. Ding, K. Lyngbaek and A. Zakhor, Automatic registration of aerial imagery with untextured 3D LiDAR models, Computer Vision and Pattern Recognition, 2008

[13] H. Hirschmüller, Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information, Computer Vision and Pattern Recognition, 2005, 807-814

[14] H. Hirschmüller, H. Mayer, G. Neukum, and HRSC CoI-Team, Stereo Processing of HRSC Mars Express Images by Semi-Global Matching, International Symposium on Geospacial Databases for Sustainable Development, 2006

[15] B. B. Mandelbrot, The Fractal Geometry of Nature, W.H. Freeman and Co., ISBN: 0716711869, New York, 1983

[16] G. Cao, X. Yang and Z. Mao, A two-stage level set evolution scheme for man-made objects detection in aerial images, IEEE Computer Vision and Pattern Recognition, 2005

[17] S. Belongie, J. Malik and J. Puzicha, Shape matching and object recognition using shape contexts, Pattern Analysis and Machine Intelligence, vol.24, no.4, 2002, 509-522

[18] H. Bay, T. Tuytelaars and L. V. Gool, SURF: Speeded Up Robust Features, Proceedings of the ninth European Conference on Computer Vision, 2006

[19] L. Liu, G. Yu, G. Wolberg and S. Zokai, Multiview Geometry for Texture Mapping 2D Images Onto 3D Range Data, IEEE Conference on Computer Vision and Pattern Recognition, vol.2, pp. 2293-2300, 2006

[20] S. Beucher., Watershed, hierarchical segmentation and waterfall algorithm. In Mathematical Morphology and its Applications to Image Processing, pages 69–76. Kluwer, 1994.

[21] Q. Wang and S. You, A Vision-based 2D-3D Registration System, IEEE Workshop on Applications of Computer Vision (WACV), Snowbird, Utah, December 7-8, 2009.