Multimodal Image Matching using Self Similarity

Jing Huang, Suya You and Jiaping Zhao
Department of Computer Science
University of Southern California
Los Angeles, CA 90089
huang10@usc.edu, suya.you@usc.edu, jiaping.zhao@usc.edu

***Abstract* - This paper presents a new image description and matching process based on internal self-similarity property of images. Various definitions of self-similarity are explored to find the best one for image matching. The method also ensures rotation and scale invariance and computational efficiency through a feature detection process. Experiments demonstrate that the proposed method increases robustness of image matching under different imaging conditions or modalities.**

I. INTRODUCTION

Most intelligent systems and decision-makings employ multiple imaging sensors and sources to create complete geospatial databases that are often structured as correlated data layers including elevation data, vector maps, 3D terrain models, and 2D imagery. These data sets have varied dimensionality, resolution, and accuracy. Such variations increase the difficulty of correlation (or match) of the data sets and the uncertainty of their interpretations. Therefore, there is immediate need of advanced techniques and methods for automatic matching and fusion of the multimodal sensing data.

Image matching is also a fundamental task in computer vision. Significant researches have been conducted in the past, especially in the area of feature based matching techniques. Some representative works include Scale Invariant Feature Transform (SIFT) [2], GLOH [3], PCA-SIFT [4], Speeded Up Robust Features (SURF) [5], Multiple View Kernel Projection (MVKP) [6] and Compact Descriptor through Invariant Kernel Projection (CDIKP) [7]. A comprehensive technical survey of feature description and matching can be found in [3].

However, most of the existing methods are designed for matching images within the same or similar imaging condition. They often fail when applied to the data with different sensor modalities. Figure 1 shows such an example of using SURF method to match an optical image and the corresponding LiDAR depth image.

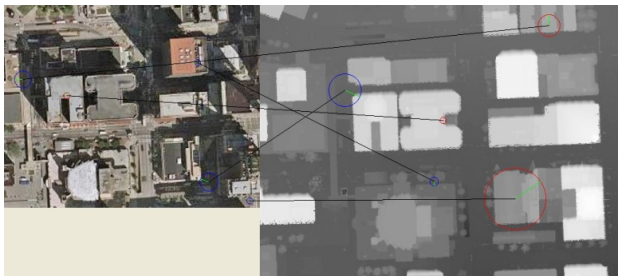

Fig. 1: Result of using SURF descriptor to match optical image and LiDAR

Our objective is to develop a new image description and matching process based on certain internal structural properties of the images. We expect the new approach to be able to capture the internal geometric layout of local regions, thus can be used to compare between images that appear substantially different at pixel level, as well as increasing the robustness of image matching under different imaging conditions and modalities.

We find that the self-similarity descriptor [1] has the potential of capturing the local internal layouts of self-similarities shared by these images, even the patterns generating those self-similarities are not shared by the images. However, there are several issues not addressed in the original framework. It lacks the rotation invariance and the scale detection process, which means we could hardly apply the approach to match images with different orientation, scale and viewpoints. In addition, the approach is computation-intensive, needing to densely compute self-similarity description (based on correlation computation) for every a few pixels. To address these problems, this paper presents a feature-based self-similarity matching approach to ensure rotation and scale invariance, as well as the computational efficiency.

II. METHODS

Figure 2 shows the framework of our approach. We first use feature extraction methods to obtain a set of distinctive image features, together with the scale and orientation estimation. Next, for each extracted feature point, we apply the correlation function combined with the scale and orientation to compute the local correlation surface. After that, the correlation surface is transformed into log-polar bins with the maximum/minimum value representing the value of each bin and then scaled such that the maximum value of all bins is equal to 1. Until now we have got the feature-based self-similarity (FB-SSIM) descriptor. Finally, we can use the Nearest Neighbor Distance Ratio Matching process to obtain the correspondences between the two images.

The remaining of this section will introduce each of the steps in detail.

A. Feature Extraction

There are several categories of feature extraction methods. The most common ones are Harris corner-based [8], Laplacian-based and Hessian-based. The representative for Laplacian-based methods is SIFT [2] with an approximation of Difference of Gaussian (DoG), while the representative for Hessian-based method is SURF [5]. For simplicity, we try the methods used in SIFT [2] and SURF [5] respectively to extract highly-distinctive local features in scale-space. The output of the feature extraction phase includes not only the position, but also

the dominant orientation and scale of the feature point. It turns out that the results from two different methods are analogous.

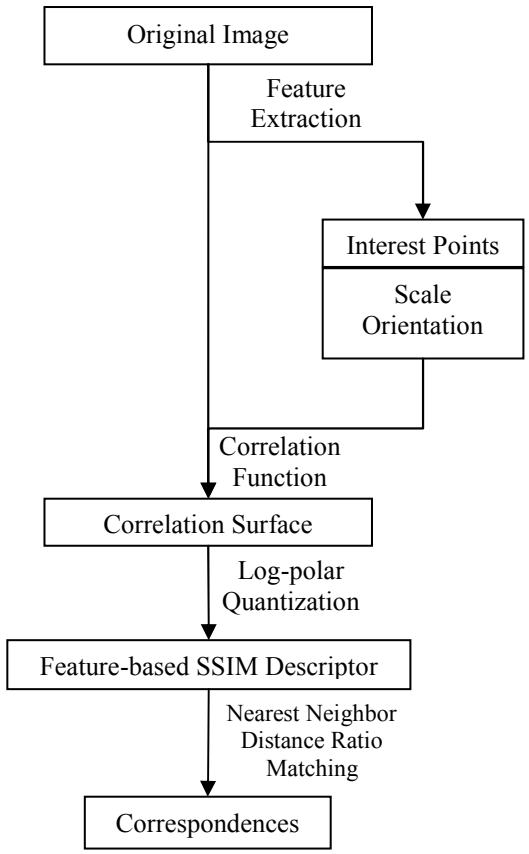

Fig. 2: Matching framework of feature-based self-similarity

B. *Correlation Surface*

The core part of self-similarity is the correlation surface, which we expect to be invariant to different sensor modalities. Although in general, the correlation surface can stretch as far as the whole image, but for better estimation of locality we typically focus on the correlation surface defined within a local region centered at a feature point q:

$$S_q(x) \coloneqq Similarity\big(P(q), P(x)\big), \forall x \in Region(q), \quad (1)$$

where P(q) is the center patch around q, and P(x) is the travelling patch decided by the position of point x in this local region of q. (See Fig. 3)

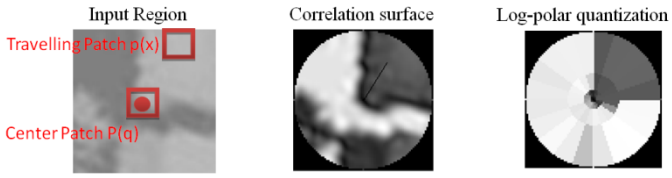

Fig. 3: From original local image region (left) to correlation surface (middle), then quantized using log-polar coordinate (right).

The definition of the correlation function in [1] is:

$$S_q(x, y) = \exp\left(-\frac{SSD_q(x,y)}{\max(var_{noise}, var_{auto}(q))}\right), \quad (2)$$

where var_{auto} is the maximum SSD result of the nearest patches to the center patch, expected to account for variance of sharpness, and var_{noise} is a constant threshold. By our notation, we can rewrite it as

$$S_q(x) = \exp\left(-\frac{SSD(P(q),P(x))}{\max(var_{noise}, var_{auto}(q))}\right). \quad (3)$$

However, in practice we find that this complicated exponential model does not always best describe the degree of similarity, in both optical images and multimodal images. We test a few correlation functions (Table 1) in our experiments. The results are obtained with densely sampled feature points and without scale and orientation changes, in order that the results are not mixed up with the feature extraction phase and thus can reflect the true performance of the correlation functions. Figure 4 shows the evaluation result of different SSIM correlation functions on multimodal data. It turns out that the simple formula (Rmse) works best in most cases.

C. *Combine Scale and Orientation with the descriptor*

To deal with the general matching problem with scale and orientation changes, we can use the detected feature scales and orientations to "normalize" each feature patch to feature's dominant scale and orientation.

In detail, let θ be the orientation at (x, y) (relative to the center point), then we assign the intensity at (x, y) of scale γ (define original scale = 1) to be:

$$I_\gamma(x, y) := I(x', y'), \quad (4)$$

where

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \gamma M(-\theta) \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \gamma cos\theta & \gamma sin\theta \\ -\gamma sin\theta & \gamma cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \quad (5)$$

Note that we can do Gaussian or bilinear interpolation to get the value of fractional coordinates if necessary.

D. *Descriptor Computing by Quantization*

Following the method described in [1], the correlation surface is transformed into a binned log-polar representation to generate a local self-similarity descriptor (Fig. 4 right). The value for each bin is represented by the maximum of the correlation function values in this bin. In our experiment there are 4 logarithmic radial bins and 12 angular bins, resulting in a descriptor of only 48 dimensions, which supports fast computation of the Euclidean distances in the matching phase without much loss of distinctiveness.

E. *Matching*

We use the Nearest Neighbor Distance Ratio (NNDR) matching with Euclidean distance as the basic metric. The best match for each feature is found by identifying its nearest neighbors (with minimum Euclidean distance) in the matched image. If we denote the nearest neighbor as N_1, and the 2nd nearest neighbor as N_2, then we say A is matched to N_1 if and only if $dist(A,N_1)/dist(A,N_2) <$ threshold. In our experiments the threshold is typically set to be between 0.6 and 0.8, and unless stated otherwise, the threshold is 0.65.

Table 1: Description of different SSIM correlation functions.

Correlation Function Name (Abbreviation)	**Mathematical Formula**	**Description**
Exponential Form (Exp)	$S_q(x) = \exp\left(-\frac{SSD(P(q),P(x))}{\max(var_{noise}, var_{auto}(q))}\right)$	The original exponential formula.
Sum of squared differences (Ssd)	$S_q(x) = \mathrm{SSD}(\mathrm{P(q)}, \mathrm{P(x)})$	Sum of squared intensity differences between the two patches.
Root mean square error (Rmse)	$S_q(x) = \mathrm{RMSE}(\mathrm{P(q)}, \mathrm{P(x)}) = \sqrt{\mathrm{SSD}(\mathrm{P(q)}, \mathrm{P(x)})}$	Square-root of the SSD surface, having the same unit as the intensity.
Average (Avg)	$S_q(x) = \lvert\mathrm{Average}(\mathrm{P(q)}) - \mathrm{Average}(\mathrm{P(x)})\rvert$	Represent the patch by the average intensity of all pixels in it and calculate the difference between them.
Normalized root mean square error (Nrmse)	$S_q(x) = \mathrm{Normalized}(\mathrm{RMSE}(\mathrm{P(q)}, \mathrm{P(x)}))$	Treat RMSE surface as a set of numbers and normalize them such that the standard deviation equals to 1.

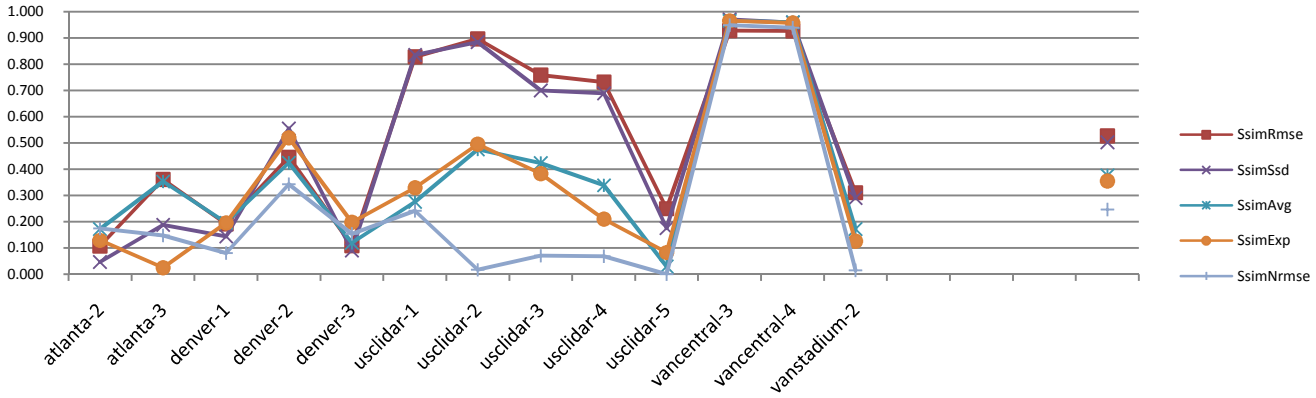

Fig. 4: Comparison of different SSIM correlation functions on multimodal data. The horizontal axis corresponds to different pairs of LiDAR intensity and depth image, while the vertical axis shows the ratio between correct matches and total matches. The threshold ratio is 0.8 in this experiment.

III. Experimental Results and Disscussion

In this section, we will present some results of image matching using the proposed self-similarity descriptor with Root-mean-square-error as the correlation function. Each correspondence is shown as a line connecting two points with the same random color in the two images respectively. For each point there's a smaller circle denoting the exact center of position and a larger circle indicating the scale on which the correlation surface is computed and quantized for this point. The white lines (bright lines) show those matches whose NNDR is below 0.6, indicating a high confidence of matching, while the gray ones indicate those matches with NNDR over 0.6 but smaller than the defined threshold, meaning relatively lower confidence of match.

A. Synthesized Data

For this part of testing, we manually adjust the intensity, or rotate/scale the original image to form the new image, and match the new image to the original one using our feature-based self-similarity descriptor. Figure 5-7 show the robustness of FB-SSIM to illumination, orientation and scale changes.

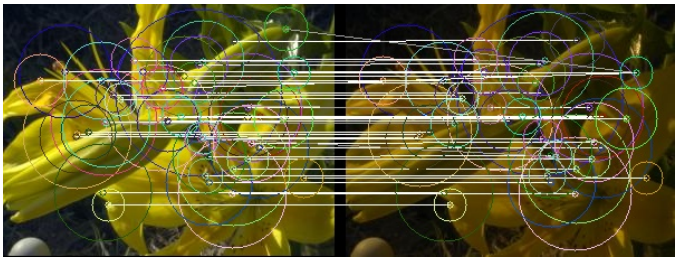

Fig. 5: Matching result between image pairs with illumination change.

B. Real Data

1) With fixed scale and orientation

In this part, we have fixed scale and orientation i.e. we assume the unified scale and orientation in the computation of descriptors.

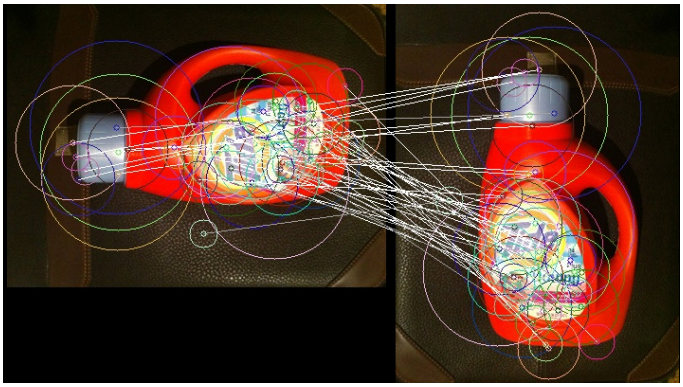

Fig. 6: Matching result between image pairs with orientation change.

Fig. 7: Matching result between image pairs with both orientation and scale changes.

The idea is that the descriptor itself is still powerful, but we have to admit that the bottlenecks in the feature-based matching on multimodal data, or even images that are not purely optical, lie not only on the descriptor, but also on the orientation, scale and even the feature extraction itself. For example, in the image pair of Fig. 8, by SIFT feature extraction the optical image can have as many as 364 feature points, while the textureless depth image contains only 20 feature points. The similar situation appears in most current feature extraction methods. Therefore, it's extremely difficult for the descriptor to proceed given so few available features with noises. Another example is given in Fig. 9. We can see that when we manually provide the knowledge that the two images share the same orientation, the self-similarity descriptor can work perfect to match the corresponding regions, however, if we use the orientation given by SIFT, SURF or simple gradient-based method for each region, there can hardly be any plausible correspondence.

Figure 14 shows the matching result between LiDAR intensity images of large area in Vancouver at different time, which demonstrate the capability of the self-similarity descriptor on large data sets.

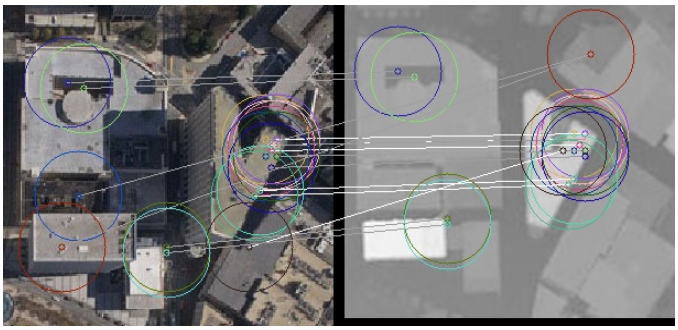

Fig. 8: Matching result between aerial image and LiDAR depth image.

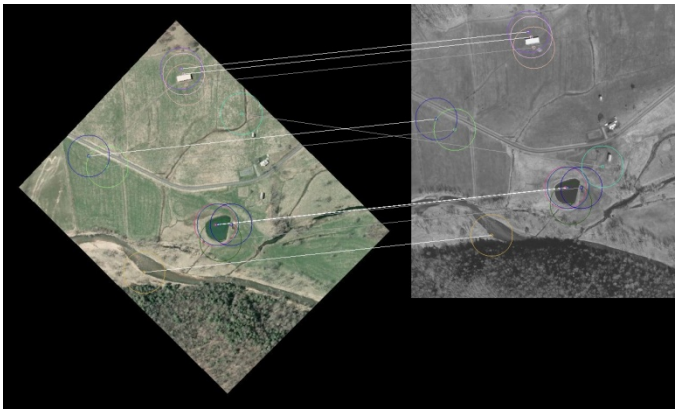

Fig. 9: Matching result between image pairs sensed differently.

2) With arbitrary scale & orientation

Although due to the absence of suitable feature extraction method in the data mentioned above, we have to manually fix the orientation and treat features densely distributed over the image, in other practical cases our feature-based self-similarity descriptor achieve favorable results. Figure 10 shows the matching result between LiDAR depth image and aerial image of the same urban area. Figure 11 shows the matching result between the low-texture surfaces. Figure 12 shows the matching result of the southernmost part of data of Figure 14. Figure 13 shows the matching result between the corresponding LiDAR intensity image and depth image.

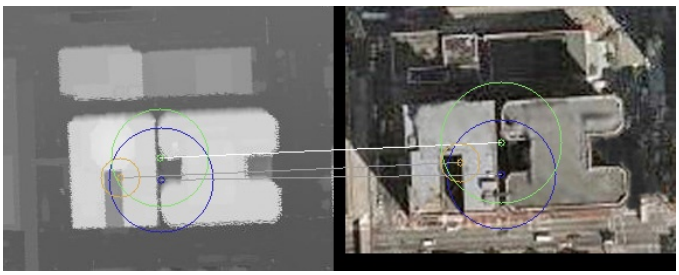

Fig. 10: Matching result between LiDAR depth image and aerial image of the same urban area.

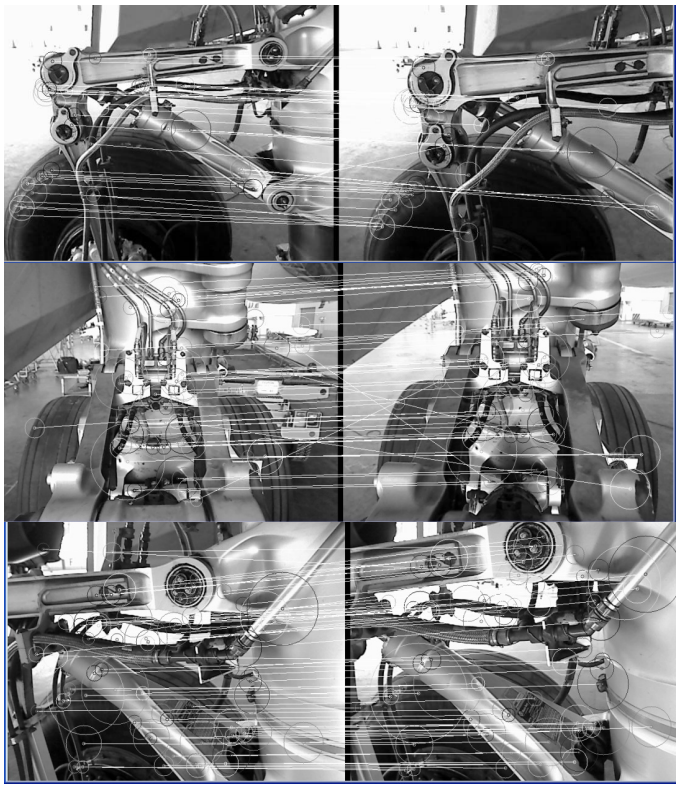

Fig. 11: Matching result between low-texture surfaces.

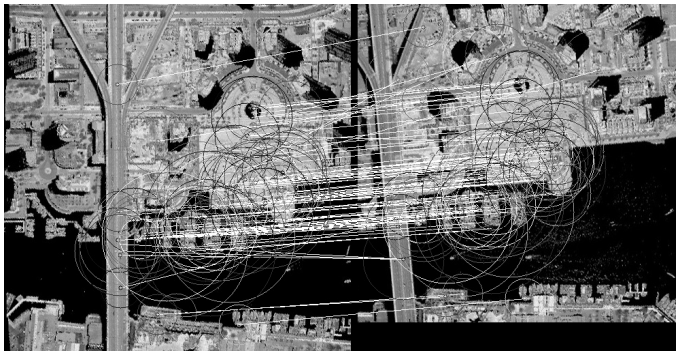

Fig. 12: Matching result between LiDAR intensity images at different time.

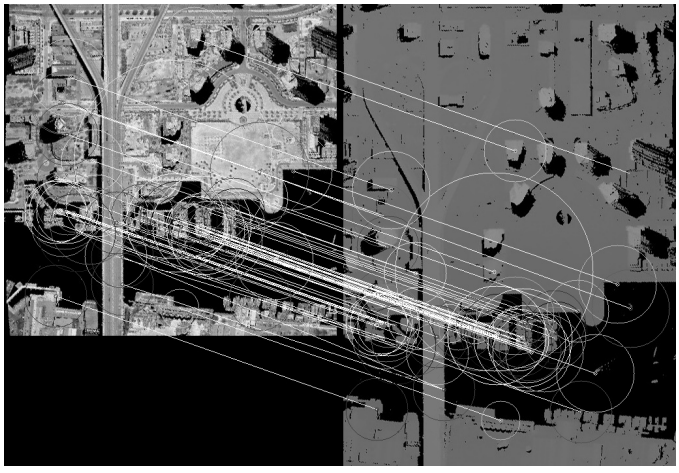

Fig. 13: Matching result between LiDAR intensity image and LiDAR depth image.

IV. Conclusion and Future Work

We explored a new image description and matching process based on image internal self-similarity. Our contributions include: (1) generalized the idea and framework of the self-similarity descriptor, especially for the multimodal data; (2) defined a "better" self-similarity function, increasing robustness of image matching under different imaging conditions; (3) extend to feature based approach, while enhancing the ability for orientation and scale invariance. The characteristic of feature-based and the low dimensionality make our matching procedure extremely fast and able to run for real-time applications. Our approach provides an alternative to the applications, such as geospatial feature matching, recognition, and fusions.

During our experiments we also find that common feature extraction methods, as well as the orientation/scale assignment, can fail in some of the challenging imaging conditions. New feature extraction and orientation assignment technique that is more invariant to modality change is needed in order to address this problem.

We would like to mention that in general, the correlation function can be pixel-based, gradient-based, edge-based or even shape-based, etc. It doesn't even require that the two self-similarity descriptors are calculated using the same function, given certain normalization procedure. As a primitive example, in Table [Function 1], the monotonicity related to similarity or dissimilarity of the exponential form is different from the other four, but it doesn't matter until when we try to match two images using different correlation functions.

In more general cases, the result of correlation function can be a conditional value, or a multi-value vector, in order to give robust estimation on a large variety of domains.

On the other hand, there can be many other forms of self-similarity. For example, the self-similarity based feature extraction [9] is also another kind of generalized usage of self-similarity.

In the future, we will continue to develop generalized self-similarity descriptors and other forms of usage, improve the feature extraction and orientation/scale measure on certain data types such as the range image, and test our approach on data from more different modalities.

Acknowledgment

This study is supported by NGA under University Research Initiatives (NURI). Some of the test data are from Airborne 1, Sanborn, and Navteq Corps. We also would like to thank the members of Computer Graphics and Immersive Technologies (CGIT) lab of USC.

References

[1] E. Shechtman and M. Irani, "Matching Local Self-Similarities across Images and Videos," in IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[2] D. G. Lowe, "Object Recognition from Local Scale-Invariant Features," in Proceedings of the 7th International Conference on Computer Vision, 1999.

[3] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 10, pp. 1615-1630, Oct. 2005.

[4] Y. Ke and R. Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors," in Proceedings of the Conference on Computer Vision and Pattern Recognition, Washington, USA, pages 511-517, 2004.

[5] H. Bay, T. Tuytelaars and L. Van Gool, "SURF: Speeded Up Robust Features," in Proceedings of the European Conference on Computer Vision, pp. 404–417, 2006.

[6] Q. Wang and S. You, "Real-time Image Matching based on Multiple View Kernel Projection," in IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[7] Y. Tsai, Q. Wang, S. You, "CDIKP: A Highly-Compact Local Feature Descriptor," IEEE International Conference on Pattern Recognition (ICPR), 2008.

[8] C. Harris, M. Stephens, "A Combined Corner and Edge Detector," in Proceedings of the Alvey Vision Conference, 1988.

[9] J. Maver, "Self-Similarity and Points of Interest," IEEE Transactions on Pattern Analysis and Machine Intelligence, vo. 32, no. 7, pp.1211-1226, July 2010.

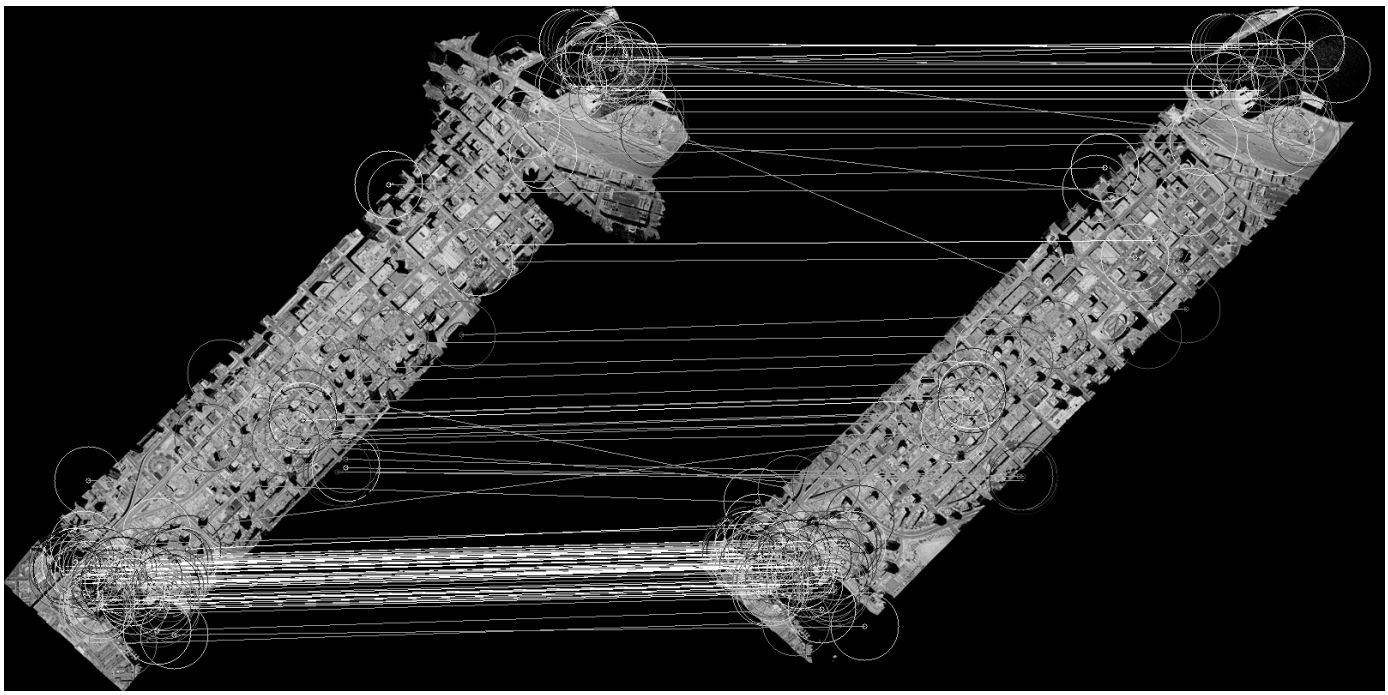

Fig. 14: Matching result between LiDAR intensity images of large area at different time.