EFFECTIVENESS OF VISUAL FEATURES ON AS-BUILT BUILDING INDOOR ENVIRONMENTS MODELING

Zhenhua Zhu
Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, Canada

ABSTRACT: *As-built information of building elements (e.g. element dimension, geometry, material, etc.) could be used to facilitate multiple building assessment and management tasks, including project progress monitoring, productivity analysis, construction inspection, etc. However, the current process for retrieving as-built information of building elements from remote sensing data is labor-intensive and time-consuming. This is especially true for modeling the building indoor environments prevalent with occlusions and partitions. In order to address these limitations, the use of RGB-D mapping has been proposed and shown a promise for modeling building indoor environments. One fundamental part in the RGB-D mapping is to select an appropriate combination of visual feature detectors and descriptors. This paper investigates the effectiveness of different visual feature detectors and descriptors on modeling 3D building scenes. Several visual feature detectors and descriptors (e.g. GFTT, SURF, SIFT, ORB, and BRISK) have been evaluated. The evaluation criteria considered in the paper include accuracy and speed. The feature detectors and descriptors have been tested in multiple building scenarios with the same hardware configuration. Based on the evaluation results, it could be found that the combination of a SURF feature detector and a BRISK feature descriptor is more accurate than the others. Meanwhile, the use of the ORB feature detector and descriptor could get the fast speed.*

KEYWORDS: *As-built building information, automation, comparative studies, RGB-D mapping*

1. INTRODUCTION

Three dimensional (3D) as-built information of building elements (e.g. columns, beams, and walls) record the actual status of buildings. Therefore, they have been identified useful for owners, designers, contractors, and facility managers in multiple building assessment and management tasks (Zhu, 2012). However, the current process of retrieving and modeling such information is labor intensive and time consuming. This is especially true in the case of building indoor environments. According to the report from a recent research study, a simple task for the generation of 3D point clouds for 40 rooms may require a 3D laser scanner to be set up at hundreds of locations due to potential indoor partitions and occlusions (Adan et al. 2011). Such labor-intensive and time-consuming nature significantly counteracts the benefits of using 3D as-built information in practice, unless the current information retrieval and modeling process can be improved, and the 3D as-built information could be updated and reviewed frequently (Pettee, 2005).

In order to reach this goal, several research studies have been proposed. Specific for the building indoor environments, the recent work built upon the RGB-D camera is promising. RGB-D stands for Red, Green, Blue plus Depth. Typically, an RGB-D camera, such as Microsoft© Kinect, is small, portable, and easy to carry, which makes it fit for the retrieval of as-built information in the building indoor environments. The camera could capture RGB-D images (i.e. pairs color and depth images simultaneously) almost in real time (30 Hz), and maintain the resolution of the images at 640x480. When an RGB-D image (i.e. a pair of the color images and depth images) is captured by the camera, a set of 3D points (i.e. point cloud) could be automatically generated. The point clouds from different RGB-D images could be further merged and aligned by being progressively mapped (i.e. RGB-D mapping).

Currently, there are many RGB-D mapping methods available (Henry et al. 2010; Engelhard et al. 2011). Their basic ideas are similar. First, the 3D point clouds are generated from the RGB-D images captured by the RGB-D camera. In the consecutive images, their visual features are detected, described, and matched. According the 2D matching results in the images, the corresponding matched 3D points in the point clouds are located. This way, the pair-wise transformation matrix between the point clouds could be estimated, and the point clouds could be registered under one 3D coordinate system. During the RGB-D mapping process, one of the critical steps lies in the

Citation: Zhu, Z. (2013). Effectiveness of visual features on as-built building indoor environments modelling. In: N. Dawood and M. Kassem (Eds.), Proceedings of the 13th International Conference on Construction Applications of Virtual Reality, 30-31 October 2013, London, UK.

selection of appropriate combinations of visual feature detectors and descriptors for the detection, description, and matching of visual features.

The objective of this paper is to evaluate the effectiveness of different combinations of the visual feature detectors and descriptors in the RGB-D mapping process. In doing so, the framework following the basic RGB-D mapping idea has been implemented. The visual feature detectors and descriptors, including Good Features to Track (GFTT) (Shi and Tomasi, 1994), Features from Accelerated Segment Test (FAST) (Rosten and Drummond, 2006), Scale-Invariant Feature Transform (SIFT) (Lowe, 2004), Speed-Up Robust Features (SURF) (Bay et al. 2008), Oriented Fast and Rotated BRIEF (ORB) (Rublee et al. 2010), etc. have been considered. The different configurations of these visual feature detectors and descriptors have been tested in multiple building scenarios. The mapping accuracy and speed have been recorded. The evaluation results indicated that the combination of a SURF feature detector and a BRISK feature descriptor (i.e. SURF/BRISK) could reach the high mapping accuracy. The use of the ORB feature detector and ORB descriptor (ORB/ORB) could get the fastest mapping speed without the support of the graphic processing unit (GPU).

2. RELATED WORK

In general, visual features refer to those local points, blobs or regions of interest in a color (RGB) image. So far, several detectors and descriptors have been developed to distinctively detect and describe the visual features, even when the color images are under certain affine deformations. Here are the details of the common ones which have been widely used in computer vision applications.

2.1 Good Features to Track (GFTT)

In 1994, Shi and Tomasi (1994) presented the concept of GFTT. They designed a GFTT detector to decide which visual features were good for the purpose of visual tracking. In their work, the strong Harris corners (Harris and Stephens, 1988) with high eigen-values were first kept. Then, in the remaining corners, those that were relatively "weak" were further rejected, if there were relatively "strong" corners close to them. Consider the corners typically appear at object boundaries where multiple motions are highly possible. Therefore, the GFTT detector was expected to address the generalized aperture problem (Senst et al. 2012), and moreover the corners kept by the GFTT detector are always those whose motions can be reliably estimated.

2.2 Features from Accelerated Segment Test (FAST)

FAST was proposed by Rosten and Drummond (2006). Similar to the GFTT, it was designed based on a corner detector. The detection procedure for the FAST included two main steps. First, the potential corner points in an image were classified with a segment test. Then, a score value was calculated at each potential corner point. The score values could be used to remove the false corners that have been classified before. In general, the FAST detector (Rosten and Drummond, 2006) has been identified as reliable and fast (Rosten et al. 2010). Therefore, it has been widely used for the applications with the real-time requirements, such as Augmented Reality workspaces (Klein and Murray, 2007).

2.3 Binary Robust Invariant Scalable Keypoints (BRISK)

BRISK was proposed by Leutenegger et al. (2011). In their framework, a scale-space pyramid was first created by progressively half-sampling an original image. The potential regions of interest on each octave and intra-octave of the pyramid were then detected with the FAST detector (Rosten and Drummond, 2006). Then, the detection results were refined with the non-maxima suppression. Moreover, the BRISK feature descriptions were provided on the detection results using the configurable circular sampling patterns. In general, the BRISK detector and descriptor could produce both distinctive, scale and rotation invariant visual features.

2.4 Scale-Invariant Feature Transform (SIFT)

SIFT was developed by Lowe (2004). In his framework, the local maxima or minima of the Difference of Gaussians (DoG) were first used to locate potential key points. Then, some of the potential key-points were removed, if they had low contrast values or were poorly localized along the edges. In the remaining key-points, their dominant orientations were assigned. When the key-points were located and assigned with dominant orientations, the feature vectors were calculated at the key-points as the feature descriptions to make the key-points highly distinctive. The SIFT key-point detector and descriptor were expected to be invariant to scale and rotation.

Also, they could be robust to illumination changes. Therefore, they have been widely used for object recognition (Sirmacek and Unsalan, 2009), robust localization and mapping in a stereo system (Se et al. 2001), panorama stitching (Brown and Lowe, 2007), etc.

2.5 Speeded-up Robust Features (SURF)

SURF was presented by Bay et al. (2008). They adopted the concept of the Hessian matrix and approximated the determinant of the Hessian matrix with two box filters (Bay et al. 2008). The size of the box filters was up-scaled. Based on the response values of an image to these two filters, the SURF key-points in the image could be located. The descriptions could be further produced as the vectors based on the distribution of the intensity content within the local image regions of the points. The SURF detector and descriptor could utilize the integral images to reduce the computation time, which makes them almost three times faster than the SIFT detector and descriptor (Bay et al. 2008). Also, similar to the SIFT detector and descriptor, they are supposed to be robust against image rotation and scale (i.e. rotation and scale invariance).

2.6 Oriented FAST and Rotated BRIEFF (ORB)

ORB is the combination of the FAST point detector (Rostern et al. 2010) and the BRIEFF descriptor (Calonder et al. 2010) with several improvements. First, Harris corner measures were calculated and used to remove potential edge points. Only those corner points with high confidence values were kept. Then, the orientation of each corner point was estimated based on the intensity centroid of the local image patch around the corner (Rublee et al. 2011). The orientation information could help to identify the corresponding BRIEFF test pattern, which made the ORB description rotation-invariant. In addition to the rotation-invariance, another benefit of using the ORB detector and descriptor is their computational efficiency. This is especially true when they are compared with the SIFT and SURF feature detectors and descriptors (Rublee et al. 2011).

3. OBJECTIVE AND SCOPE

Although several visual feature detectors and descriptors have been developed, so far, none of them is perfect. Their performances vary significantly. That is why the appropriate selection is necessary for specific computer vision applications. For example, Senst et al. (2007) compared different visual feature detectors and indicated that the FAST detector was one of the efficient feature detectors for local optical flow tracking. Chandrasekhar et al. (2010) found that the SIFT descriptor had the better performance for mobile visual search than others.

The focus of this paper has been placed on investigating the effectiveness of different combinations of visual feature detectors and descriptors in the RGB-D mapping process for retrieving and modeling as-built conditions in the building indoor environments. In order to select the appropriate combination of visual feature detectors and descriptors, this paper first implements a general framework for the RGB-D mapping. Then, the different combinations of the visual feature detectors and descriptors have been tested. Their RGB-D mapping accuracy and speed are compared. The visual feature detectors that are considered in the paper include the FAST, GFTT, SIFT, SURF, BRISK, and ORB, while the visual feature descriptors include the BRISK, SIFT, SURF, and ORB. All of these detectors and descriptors are common, and have been widely used in different computer vision applications.

4. FRAMEWORK FOR RGB-D MAPPING

A typical RGB-D mapping process includes the detection, description, and matching of visual features. Specifically, the visual features are first detected in the RGB images. These features are then distinctively described. Based on the feature descriptions, the common features in the consecutive RGB images are matched. The 2D matching results could be further extended into 3D. When the pairs of 3D matching points in consecutive point clouds are determined, the point clouds representing the building indoor environments from different RGB-D images could be merged and aligned. The overall RGB-D mapping process has been illustrated in Fig. 1.

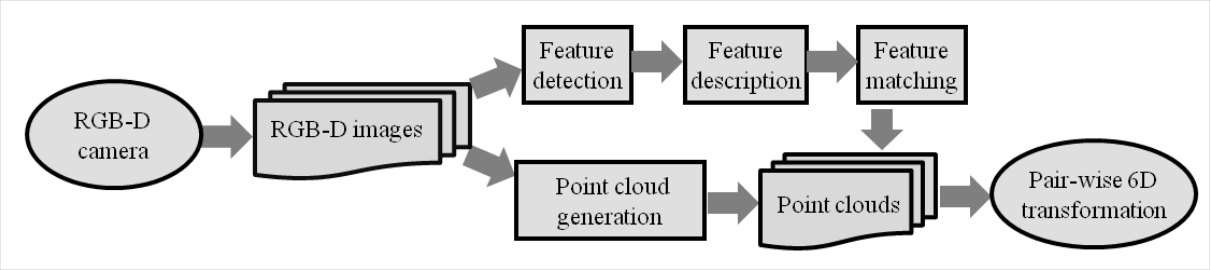

Fig. 1: Framework of RGB-D mapping.

5. EVALUATION CRITERIA AND EXPERIMENTS

The RGB-D mapping process mentioned above has been implemented and tested with different combinations of the visual feature detectors and descriptors. During the tests, the mapping accuracy and speed, as two main evaluation criteria, are recorded to measure the effectiveness and efficiency of different combinations of visual feature detectors and descriptors. Here, the mapping accuracy is determined by calculating the difference between the camera positions estimated from the mapping process (i.e. estimated trajectory) and the real positions of the camera (i.e. ground truth). The mapping speed is determined by the duration taken when performing the RGB-D mapping with each combination of the visual feature detectors and descriptors.

The specific configuration of the platform used to evaluate the RGB-D mapping process has been listed in Table 1. A total of seven RGB-D datasets have been used for the evaluation purpose. The datasets include freiburg1_xyz, freiburg1_rpy, freiburg1_360, freiburg1_desk, freiburg1_desk2, freiburg1_floor, and freiburg1_room. All were prepared by Strum et al. (2012) in the Computer Vision Groups at the Technische Universität München, each of which contains the RGB-D images plus the camera positions recorded (Strum et al. 2012). The real positions of the camera were captured by a motion capture system and they were included in the datasets to construct the ground-truth trajectories for the determination of the mapping accuracy.

Table 1: Configuration of the platform for the evaluation of the framework

Software	Operating System	Ubuntu 12.0.4 LTS
	C++ Code Compiler	gcc 4.6
Hardware	Central Processing Unit (CPU):	Intel(R) Core (TM) i7-2600K CPU @ 3.4 GHz
	Graphic Processing Unit (GPU):	NVIDIA GeForce GTX 560 Ti (1280 megabytes)
	Memory:	16 gigabytes (4x4 gigabytes)
	Motherboard	ASUS P8Z68-VPRO (Intel Z68 Chipset)
	Hard drive	Toshiba MK5061GSYN
	Operating System	Ubuntu 12.0.4 LTS (32 bits)

Different combinations of the visual feature detectors and descriptors have been implemented and tested in the RGB-D mapping process. The combinations (i.e. detector/descriptor) include BRISK/BRISK, BRISK/SIFT, BRISK/SURF, FAST/BRISK, FAST/SIFT, FAST/SURF, GHTT/BRISK, GHTT/SIFT, GHTT/SURF, ORB/BRISK, ORB/ORB,, ORB/SIFT, ORB/SURF, SIFT/BRISK, SIFT/SIFT, SIFT/SURF, SURF/BRISK, SURF/SIFT, and SURF/SURF. The implementations of these visual feature detectiors and descriptors could be found in the Open Source Computer Vision (OpenCV) library (Bradski and Kaehler, 2008). The GPU support has not been considered when implementing these feature detectors and descriptors.

Fig. 2 illustrates the results of using the combination of the SURF/SURF for the RGB-D mapping, when the dataset, freiburg1_xyz, was used for the evaluation. In the figure, it could be seen that the point clouds newly generated were progressively added into existing ones, and the number of the 3D poings kept growing. Meanwhile, the camera positions during the RGB-D mapping process could be estimated and recorded correspondingly.

Table 2 and Table 3 summarized the results for the seven datasets that have been tested so far. According to the test results, it could be found that the combination of the SURF/BRISK produced the most accurate RGB-D mapping results followed by the combination of the SURF/SIFT. However, the use of the SIFT descriptor may significantly increase the RGB-D mapping duratoin. Therefore, it is recommended to use the support of the GPU, if the SIFT descriptor has to be selected. The combination of the SIFT/SURF produced the most inaccurate RGB-D mapping results, compared with other possible combinations, and it is highly not recommended. As for the runnting time required, the combination of the ORB/ORB is taster than the other combinations. In contrast, the use of the SIFT/SIFT is the slowest.

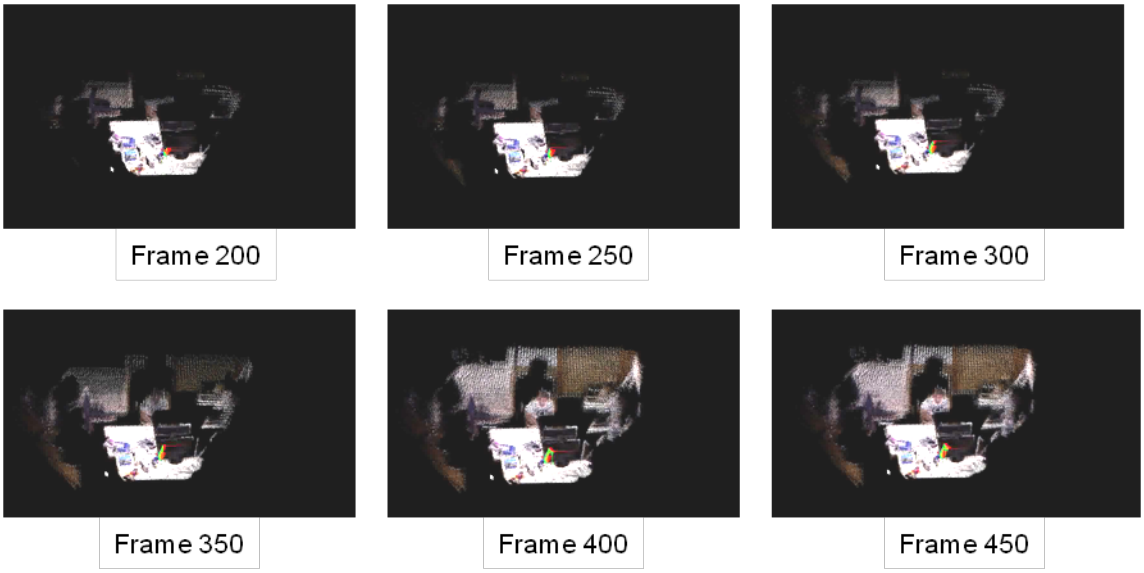

Fig. 2: RGB-D mapping results for the dataset freiburg1_xyz.

Table 2: Average error for different detector/descriptor combinations

		Detectors					
		BRISK	FAST	GFTT	SIFT	SURF	ORB
Descriptors	BRISK	0.2244	0.2690	0.1931	N/A	0.1873	0.2480
	SIFT	0.2079	0.2132	0.2272	0.3182	0.1891	0.2815
	SURF	0.1965	0.2394	0.2451	0.5420	0.2105	0.2374
	ORB	N/A	N/A	N/A	N/A	N/A	0.2402

Table 3: Running time (second) for different detector/descriptor combinations

		Detectors					
		BRISK	FAST	GFTT	SIFT	SURF	ORB
Descriptors	BRISK	10.7938	3.9642	5.3018	N/A	11.0563	2.6676
	SIFT	20.0081	3.9901	5.7614	48.7552	19.9264	8.5529
	SURF	10.6324	3.9859	5.0941	39.3383	11.0458	2.5961
	ORB	N/A	N/A	N/A	N/A	N/A	1.7230

In order to show the practicality of the proposed RGB-D mapping framework and the likelihood of uptake by the construction industry, the author used the framework for the generation of 3D point cloud of an office in the EV building at Concordia University. Considering one scan could not capture the full scene of the office, 3D points from each scan are progressively registered (Fig. 3). The 3D points generated from the proposed framework

record the building geometry and actual details of building elements, which is useful to facilitate multiple building assessment and management tasks, such as construction errors identification and on-site communication and coordination between different parties.

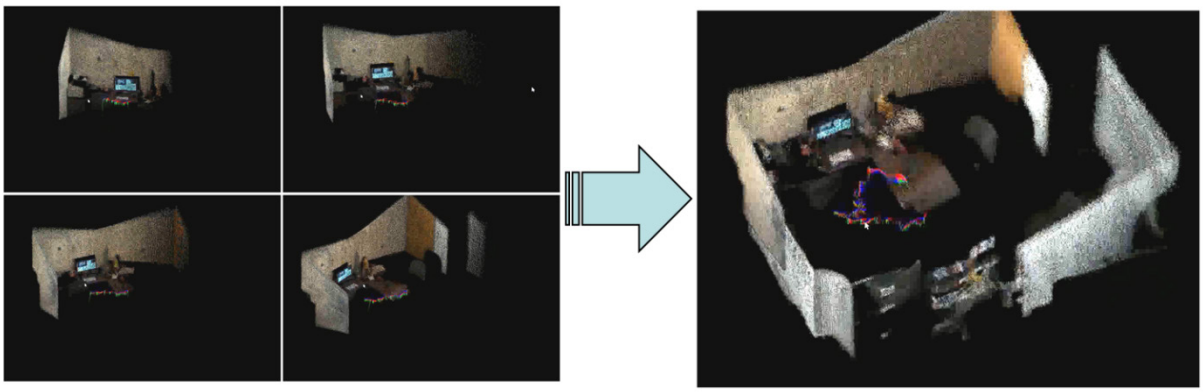

Fig. 3: 3D point clouds registration

6. CONCLUSIONS AND FUTURE WORK

3D as-built information of building elements have been identified useful, when addressing the problems related to building assessment and management. However, the current process for retrieving and modeling such information is labor intensive and time consuming. In order to address this issue, several research studies have been proposed, and the rececent work using the the RGB-D cameras is promising. The RGB-D camera is portable and convenient to use. Also, it could capture the RGB-D images and generate the 3D point clouds with the RGB-D mapping almost in real time. All of these benefits make the cammera become a good choice for the retrieval and modeling of as-built information of building elements especially in the indoor environments.

This paper investigated the effectiveness of different combinations of common visual feature detectors and descriptors on the RGB-D mapping process, considering the RGB-D mapping plays an imortant role in the retrieval and modeling of as-built information. Several common visual feature detectors (BRISK, FAST, GFTT, SIFT, SURF, and ORB) and descriptors (BRISK, SIFT, SURF, and ORB) have been selected and their different combinations have been evaluated. The main evaluation criteria include the mapping ccuracy and speed. A total of seven RGB-D datasets have been used for the evaluaiton purpose. The evaluation results from these datasets indicated that the combination of the SURF/BRISK could reach more accurate RGB-D mapping results than other possible combinations. Also, the combination of the ORB/ORB could produce the fastest registration speed, if there is no GPU support for all the combinations.

7. ACKNOWLEDGEMENT

This paper is based in part upon the work supported by the National Science and Engineering Research Council (NSERC) of Canada. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of the NSERC.

8. REFERENCES

Adan, A., Xiong, X., Akinci, B., and Huber, D. (2011). "Automatic creation of semantically rich 3D building models from laser scanner data," Proc. of ISARC 2011, Seoul, Korea

Bay, H., Tuytelaars, T., and Van Gool, L., (2008), "SURF: Speeded Up Robust Features.", Computer Vision and Image Understanding (CVIU), 110(3): 346--359.

Bradski, G. and Kaehler, A. (2008). " Learning OpenCV - Computer Vision with the OpenCV Library." O'Reilly Media Publisher, Oct. 1, 2008, ISBN-10: 0596516134

Brown, M. and Lowe, D. (2007). "Automatic panoramic image stitching using invariant features," International Journal of Computer Vision, 74(1): 59-73.

Calonder, M., Lepetit, V., Strecha, C. and Fua, P. (2010). "BRIEF: binary robust independent elementary features", In: European Conference on Computer Vision, pp. 778-792. Heraklion, Crete, Greece, Sept. 5-11, 2010

Chandrasekhar, V., Chen, D., Lin, A., Takacs, G., Tsail, S., Cheung, N-M., Reznik, Y., Grzesczuk, R., and Girod, B., (2010). "Comparison of local feature descriptors for mobile visual research." In: International Conference on Image Processing, pp. 3885-3888. Hongkong, Sept. 36-29, 2010.

Engelhard, N., Endres, F., Hess, J., Sturm, J., and Burgard, W. (2011), " Real-time 3D visual SLAM with a hand-held camera." In Proc. of the RGB-D Workshop on 3D Perception in Robotics at the European Robotics Forum, 2011.

Harris, C. and Stephens, M. (1988). "A combined corner and edge detector". In: Proc. of the 4th Alvey Vision Conference. pp. 147–151.

Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D., (2010). "RGB-D Mapping: Using Depth Cameras for Dense 3D Modeling of Indoor Environments", In: Proc. of International Symposium on Experimental Robotics, Delhi, India.

Klein, G., and Murray, D. (2007). "Parallel Tracking and Mapping for Small AR Workspaces. In: Proc. of 6th IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07). Nara, Japan, 2007.

Leutenegger, S., Chli, M. and Siegwart, R. (2011). " BRISK: Binary Robust invariant scalable keypoints." In: IEEE International Conference on Computer Vision (ICCV), pp. 2548-2555, Barcelona, Nov. 6-13, 2011

Lowe, D., (2004). "Distinctive Image Features from Scale-Invariant Keypoints", International Journal of Computer Vision, 60, 2, pp. 91-110, 2004.

Pettee, S (2005). "As-builts - problems and proposed solutions." CM eJournal, First Quarter: 1-19. <http://cmaanet.org/files/as-built.pdf> (June 10, 2012)

Rosten, E. and Drummond, T. (2006). "Machine learning for high-speed corner detection." In: European Conference on Computer Vision, Vol. 1, pp. 430–443. May, 2006.

Rosten, E., Porter, R., Drummond, T. (2010). " Faster and better: a machine learning approach to Corner Detection." IEEE Trans. PAMI 32(1): 105-119.

Se, S., Lowe, D., Little, J. (2005). "Vision-based global localization and mapping for mobile robots", IEEE Transactions on Robotics 21(3): 364-375.

Senst, T., Unger, B., Keller, I., and Sikora, T. (2012). " Performance evaluation of feature detection for optical flow tracking." In: International Conference on Pattern Recognition Applications and Methods, Vol. 2, pp. 303-309., Vilamoura, Portugal, Feb. 6-8, 2012

Shi, J. and Tomasi, C. (1994). "Good features to track.", In: Computer Vision and Pattern Recognition, pp. 593 - 600. Seattle, WA., Jun 21 - 23, 1994.

Sirmacek, B. and Unsalan, C. (2009). "Urban area and building detection using SIFT keypoints and graph theory" IEEE Transactions on Geoscience and Remote Sensing, 47(4): 1156-1167.

Sturm J., Engelhard, N., Endres, F., Burgard, W., and Cremers, D. (2012). "A Benchmark for the Evaluation of RGB-D SLAM Systems", In: Proc. of the International Conference on Intelligent Robot Systems (IROS), Oct. 2012.

Zhu, Z. (2012) "Automated As-built Modeling with Spatial and Visual Data Fusion." Proc. of 12th International Conference on Construction Applications of Virtual Reality 1-2 November, 2012, Taipei, Taiwan