Multi-View Matching for Onsite Construction Resources with Combinatorial Optimization

B. Zhang[a], Z. Zhu[a], A. Hammad[b] and W. Aly[c]

[a]Department of Building, Civil, and Environmental Engineering, Concordia University, Canada
[b] Concordia Institute for Information Systems Engineering, Concordia University, Canada
[c] Indus.ai Inc., Canada

E-mail: bingfeizh@gmail.com, zhenhua.zhu@concordia.ca, hammad@ciise.concordia.ca, walid.aly@indus.ai

Abstract

When multiple video cameras are set up to monitor construction activities, onsite construction resources (e.g. equipment and workers) might be captured by two or more cameras at the same time. It becomes important to identify whether the resources captured into separate camera views refer to the same one on the site. Otherwise, the automatic reporting of onsite resources utilization will produce repetive countings. This paper proposes a novel method for matching onsite construction resources in multiple camera views. The method relies on the visual features of a construction site and the spatial relationships of the resources on the site as matching cues. It starts with searching potential matching candidates between the camera views following their epipolar constraints. Then, the candidates' local triangular coordinates are calculated to define matching costs. This way, the matching of multiple construction resources between camera views could be solved through combinatorial optimization. The proposed method has been tested to match workers, equipment and traffic cones within the images and videos captured from construction sites. The test results showed that the method could reach an average of 93% matching accuracy.

Keywords

Onsite resources matching; Epipolar geometry; Combinatorial optimization; Multiple camera views

1 Introduction

The recent fast development of digital camera technology made it possible to set up multiple cameras to monitor construction sites [1]. These cameras can capture the detailed information on the utilization of construction resources (e.g. equipment, workforce, and materials), which could help to facilitate several construction management tasks [1]. For example, the working states of construction equipment were analyzed to estimate its productivity [2]. Also, equipment poses could be extracted to improve onsite safety in equipment operations [3].

When there are multiple cameras on a construction site, they typically have overlapping fields of view (FOVs). It is important to match the visual appearances of construction resources captured in the overlapping FOVs to find out which visual appearances refer to the same construction resource onsite. The successful matching is expected to reduce the repetitive counting and identification of construction resources. Also, it is one of essential steps to locate the resources on the site from the triangulation [4]. Moreover, if one resource is heavily occluded in one camera view, it could still be detected and tracked, as long as its occlusion in another camera view is not severe.

So far, several research studies have been proposed to match the visual appearances of generic objects of interest under multiple camera views. For example, Hu et al. [5] relied on a set of feature points through the Scale-Invariant Feature Transform (SIFT) [6, 7], Speeded Up Robust Features (SURF) [8], etc., to describe and match the object visual appearances in different camera views. Cai and Aggarwal [9] facilitated the object matching by referring to the epipolar constraint. According to the epipolar constraint, it was indicated that the projection of an object point in one camera view could generate a line (i.e. the epipolar line) in another camera view on which its corresponding projection must lie [9]. This way, the searching space is narrowed down to a line.

Most of existing generic matching methods have a limited use on construction sites. This is partly because construction video cameras are typicall set up at heights with wide camera baselines and large differences in view orientation. As a result, the construction resources in each camera view appear small with large variations. It could be difficult to find sufficient common object visual feature points for the matching purpose. On the

other hand, the use of the epipolar line could not match the resources one to one, although it does help to limit the matching search space. When the visual appearances of multiple similar construction resources lie along the same epipolar line, the matching fails with the sole use of epipolar constraints.

The goal of this paper is to address these limitations by proposing a novel method for matching the visual appearances of construction resources between camera views. The method is built upon the visual features on a construction site, as well as the spatial relationships of the construction resources on the site. Specifically, potential matching candidates between camera views are first found following the epipolar constraints. Then, the local triangular coordinates of the candidates in each camera view are calculated. It is assumed that the same resource in different camera views should have the same local triangular coordinates. This way, the matching of multiple construction resources between two camera views could be solved by minimizing the matching cost through the combinatorial optimization.

The effectiveness of the proposed method has been tested with the images collected from a real construction site. The tests were conducted under different weather and illumination conditions. According to the test results, it was found that the overall matching accuracy could reach 93% (93% for construction workers, 100% for excavators, and 92% for traffic cones separately). Moreover, the proposed method was compared with the research work of Lee et al. [4]. The comparison results indicated that the proposed method could successfully match small-sized construction resources even if their visual appearances in one camera view lie on the same epipolar line.

2 Related Work

So far, numerous generic object matching methods have been created. These methods could be classified into two categories based on their matching cues. The methods in the first category relied on the visual features of the objects in each camera view; and the methods in the second category relied on the spatial relationships of the objects under the camera views.

2.1 Matching with Visual Features

Under the matching with visual features, the visual appearances of an object under different camera views are first characterized by a set of local point or area features [10]. Then, the matching is directly conducted by checking whether the visual appearances in camera views have the same local point or area features. If the same point or area features are found, it indicates that these visual appearances refer to the same object. Otherwise, they belong to different objects.

Examples of point feature detectors and descriptors include SIFT [6, 7], SURF [8], etc. The SIFT features are typically robust to the orientation changes of camera views; however, they might be sparse for matching object visual appearances between camera views. The SURF features could be detected in a faster way; but they are not fully affine invariant [11].

Compared with point features, the area features typically refer to the visual patterns in small, local image windows. Those area-feature based matching methods first find seed points and propagate from these points into small image windows. Then, the cross-correlation of the visual patterns in these windows is conducted for the matching purpose. For example, Pratt [12] used the local image intensities for the cross-correlation; and Rashidi et al. [13] adopted the adaptive color difference.

In general, the area-feature based matching methods could produce dense matching results [14] and are robust to local affine distortions. However, the matching might still fail. This is especially true when distinctive visual patterns could not be found in the local image windows, or when the patterns experienced significant deformations from the image transformations [15].

2.2 Matching with Spatial Relationships

The epipolar geometry is one of the common common spatial relationships that have been investigated in the matching procedure. In the epipolar geometry, if the projection of a three-dimensional (3D) point X on the left view is known, its epipolar line on the right view could be calculated. Morover, the projection of the point X on the right view must be on the line. Such spatial contratint signficnatly reduces the search space in the matching process [16].

Zhang et al. [17] used the Least Median of Squares (LMedS) to find the epipolar geometry between two camera views. The method of Lee et al. [4] then relied on the epipolar geometry to match onsite construction workers captured in two camera views by considering the locations of the workers in the camera views as well as their distances to the corresponding epipoar lines. Their matching recall and precision could reach 71.4% and 98.7%, respectively [4]. Recently, Konstantinou and Brilakis [18] proposed an idea that combined the epipolar geometry with the shift of the workers' centroids and visual features across video frames to improve the matching accuracy.

2.3 Gaps in Body of Knowledge

There are several limitaitons, when adopting the existing methods to match construction resources. First, the video cameras are typically set up at height on construction sites. Their shooting distance to the onsite

construction resources (mobile equipment, workers, etc.) is long. As a result, the resources captured in the camera views always appear small, which makes it difficult to find effective point or area visual features to characterize them.

Second, most of existing matching methods based on visual features failed to match construction resources, when they have similar visual appearances. For example, all taffic cones look similar. Therefore, methods based on visual features for matching traffic cones between camera views have errors .

In addition, the matching methods based on spatial relationship might also fail due to the errors introduced in the calculation of the epipolar lines. For example, both methods proposed by Lee et al. [4] and Konstaninou and Brilakis [18] assumed the centroid of the bounding box of a worker as his/her location in one camera view, and determined the corresponding epipolar line in another view. When the worker is partially occluded in the first view, the centroid of the bounding box does not reflect his or her accurate location. As a result, the epipolar line in the second view is deviated to another worker instead, and the matching error is produced.

3 Objective and Proposed Methodology

The main objective of this paper is to propose a novel method for automatically matching onsite construction resources (e.g. mobile equipment, worker, and temporary facility) captured into camera views. The method is expected to be robust to the changes due to dynamic site activities as well as different illumination and weathe conditions. Also, it is common that construction onsite resources experience occlusions in the camera views. The proposed method is required to be able to match the resources even if they are partially occluded.

This paper focuses on matching excavators, workers, and traffic cones on a construction site. They represent three types of onsite resources, i.e. mobile equipment, labor, and temporary facilities. These resources of interest are first identified through the visual detection and/or tracking. However, many detection and tracking methods have been developed in the past, which could be used in this research.. This paper selected the Single Shot multi-box Detection (SSD) detection method [23] and the Kernelized Correlation Filters (KCF) tracking method [24].. However, other methods could be applied as well.

There are two main steps in the proposed method. The first step is to detect and match the visual feature points of the overall construction site under different camera views. The visual feature points help to establish the epipolar geometry within each pair of camera views, so that the potential matching candidates could be found. The second step is to generate a dynamic triangular mesh in each camera view with the visual feature points of the site. The triangular coordinates of the potential matching candidates in the corresponding meshes are calculated. It is assumed that the same resource in different camera views should have the same local triangular coordinates. Therefore, the difference in the triangular coordinates is defined as the matching cost. The successful matching of multiple onsite resources in different camera views should find a minimum sum of matching costs through combinatorial optimization.

The overall framework of the proposed method is illustrated in Figure 1. The following two sections introduce the details of the steps in the framework. It is worth noting that the matching method proposed in this paper does not require that the cameras are of the same type.

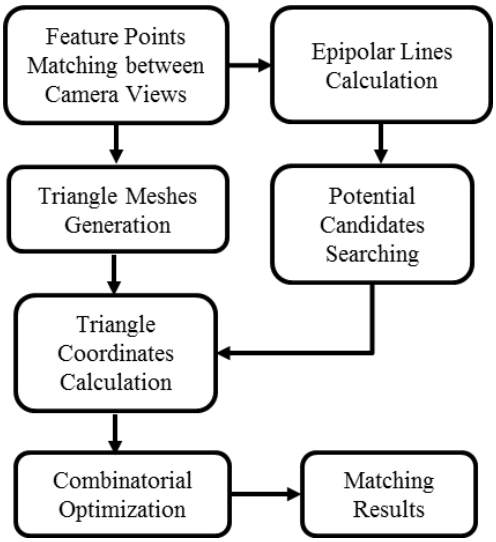

Figure 1: Proposed framework

3.1 Search for Potential Candidates

The purpose of the search for matching candidates is to reduce the matching space and improve the potential matching accuracy. The work here is similar to the previous research study proposed by Lee et al. [4]. Suppose there are two camera views, i.e. CamView1 and CamView2. An initial set of matched feature points in both camera views is first identified with the SIFT detector/descriptor [6, 7]. The selection of this detector/descriptor is mainly due to its robustness to large perspective and/or scale changes [25]. The matched feature points are further refined with the RANdom Sample Consensus (RANSAC) method [26], according to the suggestions from Hartley and Zisserman [27].

The refined feature points are used to generate a 3×3 fundamental matrix. The matrix indicates the hidden epipolar constraint between the two camera views. This way, for each resource of interest in CamView1, its corresponding epipolar line in CamView2 could be determined. The distances of the resources of interest in CamView2 to the epipolar line are calculated. Only those resources with distances equal to or smaller than their size are kept as the potential candidates for matching.

3.2 Pairwise Matching with Combinatorial Optimization

When the potential candaidates are found, the proposed method tries to address the matching problem between camera views with combinatorial optimization. Specifically, suppose there are *n* resources of interest {O_1, O2, O3, …, O_n} identified in CamView1, and their *m* matching candidates in CamView2 from the previous step are {C1, C2, C3, …, Cm}. Then, an *n* by *m* matrix, *M*, can be formulated as shown in Eq. 1.

$$M = \begin{bmatrix} M_{11} & \cdots & M_{1m} \\ \vdots & \ddots & \vdots \\ M_{n1} & \cdots & M_{nm} \end{bmatrix} \quad (1)$$

Where the element, M_{ij}, in the matrix indicates the matching cost, if the i^{th} resource (Oi) in CamView1 is assumed to match with the j^{th} candidate (Cj) in CamView2. The specific matching cost, *Mij*, is calculated as follows. First, the Delaunay triangulation process [28] is applied upon the correctly matched feature points in CamView1 to generate a triangular mesh (TM1). The mesh (TM1) is further projected to CamView2 to form another triangular mesh (TM2). The local triangle coordinates of each resource in TM1 and each candidate in TM2 are calculated. As for the ith resource (Oi) in CamView1 and the jth candidate (Cj) in CamView2, their matching cost is defined as the difference in terms of their triangle coordinates, as shown in Eq. 2.

$$M_{ij} = \sqrt{(O_{i1} - C_{j1})^2 + (O_{i2} - C_{j2})^2 + (O_{i3} - C_{j3})^2} \quad (2)$$

Where $\{O_{i1}, O_{i2}, O_{i3}\}$ and $\{C_{j1}, C_{j2}, C_{j3}\}$ are the triangle coordinates of the resource (Oi) and the candidate (Cj) in TM1 and TM2 separately. If the candidate (Cj) is not in the list of potetnials candidates for the resource (Oi), their matching cost is then set as $+\infty$.

When the matching costs in the matrix, *M*, are all determined, the matching between the resources in CamView1 and the candidates in CamView2 is transformed into an assignment problem. Here, the Hungarian algorithm [28] is adopted to find the best assignment with the total minimum matching costs. It is worth noting that *M* does not have to be a square matrix. The matching could still be made when the number of the objects of interest in CamView1 is not the same as the number of the candidates in CamView2.

4 Implementation and Results

4.1 Implementation

The matching method proposed in this paper has been implemented in Python platform under the support of the OpenCV and Munkres libraries [29, 30]. It was tested on a Microsoft Windows 10 64-bit operating system. The hardware configuration for the tests includes an Intel® Core™ i7-7700HQ CPU (Central Processing Unit) @ 2.80 GHz, a 16 GB memory, and an NVIDIA GeForce GTX 1070 GDDR5 @ 8.0 GB GPU (Graphic Processing Unit).

4.2 Results

The images and videos from real construction sites in Canada were used for the tests. One example is illustrated in Figure 2. On the site, four high definition video cameras were placed to record daily construction activities for a period of 6 months (August, 2015 ~ Feburary, 2016). 51 videos with the total size of 1.3 TBs were collected for analysis. These videos were captured under different environmental conditions (e.g. daytime vs. nighttime; and sunny vs. rainy vs. snowy).

Figure 2: Example of the sites for tests

The videos were used to test the tests of matching workers, excavators and traffic cones. Figure 3 shows the examples of the matching results from the tests. It can be seen that the workers, excavators and traffic cones could be matched between camera views. Also, the matching of the resources is possible even when they experienced occlusions. For example, one excavator is partially occluded by another one in one camera view. The method could still identify the matched excavators in the other camera view. Figure 4 shows the matching examples under different

environmental conditions.

Figure 3: Examples of matching workers, excavators, and traffic cones

Figure 4: Examples of testing under different environmental conditons

All matching results have been compiled in Table 1. It could be seen that the average matching accuracy of the proposed method is 93% (construction workers: 93%, excavators: 100%, and traffic cones: 92%). Also, the matching proposed method was compared with the research work of Lee et al. [4]. As shown in Figure 5, the work of Lee et al. [4] could not always successfully match construction resources when they are close to the same epipolar lines and/or partially occluded. The method proposed in this paper addressed such challenges well.

Table 1: Summary of matching results

	Correct Pairs	Total Pairs	Accuracy
Workers	213	229	93%
Excavators	40	40	100%
Traffic Cones	100	109	92%

Total	353	378	94%

Figure 5: Comparison between Lee et al. [4] (top) and the proposed method (bottom)

4.3 Discussion

According to the test results, it was noted that environmental conditions showed little impacts on the effectiveness of the proposed method. In most cases, the matching accuracy could reach above 90%.

On the other hand, the matching accuracy of the proposed method might be reduced when the size of the resources of interest becomes small. For example, the size of the workers or traffic cones is smaller than the size of the excavators in the tests. That might explain why the matching accuracy for excavators was 100%, while the accuracy rates for matching workers and traffic cones are 93% and 92%, respectively. In addition, the workers and traffic cones could be close to each other in the tests, which also made the matching more challenging.

Another important factor influencing the matching accuracy of the proposed method lies in the camera setup conditions (e.g. the shooting angle between two camera views). The number of common visual feature points detected under each view is reduced when both cameras have a wide view angle. Few feature points will affect the accuracy of the fundamental matrix as well as the generation of the triangular meshes.

5 Conclusions and Future Work

The accurate and robust matching of onsite construction resources of interest under different camera views is still challenging. This paper proposed a novel matching method, which solve the matching problem with combinatorial optimization. Compared with the previous work, the method works even when the resources are close to each other. The method has been tested with the images and videos collected from a real construction site in Canada. The matching was conducted under different lighting and weather conditions. The test results showed that the accuracy rates for matching construction workers, excavators, and traffic cones were 93%, 100%, and 92%, respectively. Overall, the matching accuracy could reach 93%. This paper did not evaluate the relationship between the matching accuracy and the onsite camera setups. Therefore, in the future, the focus will be placed on studying the impacts of the camera setups (e.g. camera distance and shooting angles) on the matching accuracy.

Acknowledgement

This research was funded by the Mitacs Accelerate Program via Grant IT08504. The authors gratefully acknowledge Mitacs's support. Any opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of Mitacs.

References

[1] Katz, I., Saidi, K., & Lytle, A. (2008). The role of camera networks in construction automation. In: Proc. of 2008 International Symposium on Automation and Robotics in Construction (ISARC), Vilnius, Lithuania. DOI: 10.22260/ISARC2008/0049

[2] Bügler, M., Borrmann, A., Ogunmakin, G., Vela, P. A., & Teizer, J. (2017). Fusion of photogrammetry and video analysis for productivity assessment of earthwork processes. Computer-Aided Civil and Infrastructure Engineering, 32(2), 107-123. DOI: 10.1111/mice.12235

[3] Soltani, M., Zhu, Z., and Hammad, A. (2017). Skeleton estimation of excavator by detecting its parts. Automation in Construction, 82 (2017): 1-15. DOI: 10.1016/j.autcon.2017.06.023

[4] Lee, Y. J., Park, M. W., & Brilakis, I. (2016, January). Entity Matching across Stereo Cameras for Tracking Construction Workers. In Proceedings of the International Symposium on Automation and Robotics in Construction. DOI: 10.22260/ISARC2016/0081

[5] Hu, X., Tang, Y., & Zhang, Z. (2008). Video object matching based on SIFT algorithm. In Neural Networks and Signal Processing, 2008 International Conference on (pp. 412-415). IEEE. DOI: 10.1109/ICNNSP.2008.4590383

[6] Lowe, D. G. (1999). Object recognition from local scale-invariant features. The proceedings of the seventh IEEE international conference on

Computer vision (Vol. 2, pp. 1150-1157). IEEE. DOI: 10.1109/ICCV.1999.790410

[7] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110. DOI: 10.1023/B:VISI.0000029664.99615.94

[8] Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. Computer Vision– European Conference on Computer Vision 2006, 404-417. DOI: 10.1007/11744023_32

[9] Cai, Q., & Aggarwal, J. K. (1996, August). Tracking human motion using multiple cameras. Proceedings of the 13th International Conference on Pattern Recognition (Vol. 3, pp. 68-72). IEEE. DOI: 10.1109/ICPR.1996.546796

[10] Wu, B., Zhang, Y., & Zhu, Q. (2011). A triangulation-based hierarchical image matching method for wide-baseline images. Photogrammetric Engineering & Remote Sensing, 77(7), 695-708. DOI: 10.14358/PERS.77.7.695

[11] Pang Y, Li W, Yuan Y, et al. Fully affine invariant SURF for image matching. Neurocomputing, 2012, 85: 6-10. DOI: 10.1016/j.neucom.2011.12.006

[12] Pratt, W. K. (1991). Digital Image Processing, 2nd ed., Wiley, New York. ISBN-13: 978-0471767770

[13] Rashidi, A., Fathi, H., & Brilakis, I. (2011). Innovative stereo vision-based approach to generate dense depth map of transportation infrastructure. Transportation Research Record: Journal of the Transportation Research Board, (2215), 93-99. DOI: 10.3141/2215-10

[14] Lhuillier, M., & Quan, L. (2002). Match propagation for image-based modeling and rendering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(8), 1140-1146. DOI: 10.1109/TPAMI.2002.1023810

[15] Joglekar, J., & Gedam, S. S. (2012). Area based image matching methods—A survey. Int. J. Emerg. Technol. Adv. Eng, 2(1), 130-136.

[16] Papadimitriou, D. V., & Dennis, T. J. (1996). Epipolar line estimation and rectification for stereo image pairs. IEEE transactions on image processing, 5(4), 672-676.

[17] Zhang, Z., Deriche, R., Faugeras, O., & Luong, Q. T. (1995). A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. Artificial intelligence, 78(1-2), 87-119.

[18] Konstantinou, E., & Brilakis, I. (2015). 3D Matching of Resource Vision Tracking Trajectories. In Construction Research Congress 2016 (pp. 1742-1752).

[19] Jonker, R., & Volgenant, T. (1986). Improving the Hungarian assignment algorithm. Operations Research Letters, 5(4), 171-175. DOI: 10.1016/0167-6377(86)90073-8

[20] Bertsekas, D. P. (1981). A new algorithm for the assignment problem. Mathematical Programming, 21(1), 152-171. DOI: 10.1007/BF01584237

[21] Goldberg, A. V., & Kennedy, R. (1995). An efficient cost scaling algorithm for the assignment problem. Mathematical Programming, 71(2), 153-177. DOI: 10.1007/BF01585996

[22] Chaobo, Y., & Qianchuan, Z. (2008, July). Advances in assignment problem and comparison of algorithms. In Control Conference, 2008. CCC 2008. 27th Chinese (pp. 607-611). IEEE. DOI: 10.1109/CHICC.2008.4605832

[23] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

[24] Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2015). High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(3), 583-596. DOI: 10.1109/TPAMI.2014.2345390

[25] Zhu, Q., Wu, B., & Tian, Y. (2007). Propagation strategies for stereo image matching based on the dynamic triangle constraint. ISPRS Journal of Photogrammetry and Remote Sensing, 62(4), 295-308. DOI: 10.1016/j.isprsjprs.2007.05.010

[26] Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6), 381-395.

[27] Hartley, R., & Zisserman, A. (2003). Multiple view geometry in computer vision. Cambridge university press. ISBN-13: 978-0521540513

[28] Lee, D. T., & Schachter, B. J. (1980). Two algorithms for constructing a Delaunay triangulation. International Journal of Computer & Information Sciences, 9(3), 219-242. DOI: 10.1007/BF00977785

[29] Beyeler, M. (2015). "OpenCV with Python Blueprints." Packt Publishing, ISBN-13: 978-1785282690

[30] Pilgrim, R. (2017) "Munkres' Assignment Algorithm", CSC 445 Readings <http://csclab.murraystate.edu/~bob.pilgrim/445/munkres.html> (August 3, 2017)