Automation in Construction 91 (2018) 206–215

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Automatic matching of construction onsite resources under camera views

Bingfei Zhang[a], Zhenhua Zhu[a,*], Amin Hammad[b], Walid Aly[c]

[a] Department of Building, Civil, and Environmental Engineering, Concordia University, Montreal H3G 1M8, Canada
[b] Concordia Institute for Information Systems Engineering, Concordia University, Montreal H3G 1M8, Canada
[c] GreenOwl Mobile, 57 Glen Cameron Road, Markham, ON L3T 1P3, Canada

ARTICLE INFO

Keywords:
Automatic matching
Construction onsite resources
Camera views
Combinatorial optimization

ABSTRACT

When a video camera network is placed on a construction site to monitor onsite activities, construction resources, such as equipment and worker, might be captured by two or more cameras at the same time. Therefore, it is important to conduct the matching to identify whether the resources captured into separate camera views refer to the same one on the site. Otherwise, it leads to the repetitive counting, when analyzing the onsite resources utilization automatically. This paper proposes a novel matching method that relies on the construction site visual features and the spatial relationships of onsite construction resources as the matching cues. Specifically, the method first searches the potential matching candidates between two camera views following their epipolar constraints. Then, the triangular coordinates of these candidates are calculated based on their locations in the triangular mesh of each camera view. This way, the matching of multiple construction resources between two camera views could be converted to a combinatorial optimization problem and solved with the Hungarian algorithm. The proposed method has been tested with the images and videos captured from real construction sites. The test results showed that the average matching accuracy could reach 93%.

1. Introduction

It has become common to set up a video camera network on a construction site to monitor the working environments due to the recent fast development of digital camera technology [1]. The onsite surveillance cameras in the network capture the detailed construction resources (e.g. equipment, workforce, and materials) and their related construction activity information into time-lapse images and/or videos, which could be used to facilitate multiple construction management tasks [1]. One example is to detect and track the construction equipment that was captured by the cameras to analyze its activities and to estimate the corresponding construction productivity [2]. Moreover, the equipment pose could be estimated to increase construction safety during the equipment operations [3].

When multiple cameras are placed on a construction site as a camera network, they might have overlapping field of views (FOVs). The construction resources in the overlapping FOVs can be captured by two or more cameras at the same moment. Their visual appearance in each camera view varies. Therefore, it is important to match these visual appearances in the camera views to figure out which visual appearances refer to the same construction resource on the site, in order to remove the repetitive counting and identification. Also, the successful matching of the visual appearances from the same construction resources in different camera views could bring other benefits. For example, it is one of essential steps to conduct the triangulation and determine the resource's three-dimensional (3D) location on the site [4]. In addition, 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, there are several research studies proposed for matching the visual appearances of generic objects of interest under different camera views. For example, Hu et al. [5] first described the object visual appearance under each camera view with a set of feature points through the Scale-Invariant Feature Transform (SIFT) [6,7], Speeded Up Robust Features (SURF) [8], etc. Then, the matching of the object visual appearances between the camera views was conducted by finding their common visual feature points. Also, Cai and Aggarwal [9] investigated the use of the epipolar constraint to facilitate the matching of the object projections. The epipolar constraint 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 space for searching the matched object visual appearances between two camera views could be narrowed down.

However, existing matching methods have limitations, when being adopted in real construction sites. For example, construction video cameras are set up at heights with wide camera baselines and large

* Corresponding author.
E-mail addresses: zhenhua.zhu@concordia.ca (Z. Zhu), hammad@ciise.concordia.ca (A. Hammad), walid.aly@indus.ai (W. Aly).

https://doi.org/10.1016/j.autcon.2018.03.011
Received 2 September 2017; Received in revised form 28 February 2018; Accepted 3 March 2018

differences in view orientation, the visual appearance of one construction resource in each camera view is small and different from its appearance in the views of other cameras. As a result, it is difficult to find enough common visual feature points on construction resources to conduct the matching. On the other hand, the use of epipolar line does help to limit the matching search space, but it could not match the resources one on one, especially when the visual appearances of multiple similar construction resources lie along a same epipolar line.

In order to address these limitations, this paper proposes a novel method for matching the visual appearances of construction resources captured in different camera views. The method utilized the visual features on an overall construction site as well as the spatial relationships of the onsite construction resources. It consists of two main steps. First, the method finds the potential matching candidates between two camera views following the epipolar constraints. Then, a dynamic triangular mesh in each camera view is generated. The triangular coordinates of the candidates are calculated based on their locations in the corresponding triangular mesh. The coordinate difference between each pair of potential matching candidates is defined as their matching cost. This way, the matching of multiple construction resources between two camera views could be solved by finding the minimum matching cost through the combinatorial optimization.

The method has been tested with the images collected from a real construction site under different environmental (e.g. weather and illumination) conditions. The effectiveness of the method on matching construction workers, excavators and traffic cones has been evaluated. The test results showed that the accuracy for matching construction workers, excavators, and traffic cones could reach 93%, 100%, and 92%, respectively. The overall matching accuracy with the proposed method is 93%. Also, compared with the previous research work proposed by Lee et al. [4], the method could successfully match small construction resources even if their visual appearances in one camera view lie in a same epipolar line.

2. Related work

The matching of generic objects under different camera views is a challenging task. It is especially true for construction resources, considering that construction sites are typically complex, cluttered, and large-scale. So far, numerous methods have been created to improve the matching accuracy and robustness. These matching methods could be generally classified into two categories based on the matching features they adopted. The methods in the first category relied on the object visual features in each camera view as the matching cues, while the methods in the second category focused more on the spatial relationships of the objects.

2.1. Visual feature-based matching

The point- and area- features are commonly adopted for object matching between two camera views [10]. Specifically, the visual appearances of an object under different camera views are first characterized by a set of local point or area features. Then, the visual appearances in two camera views are assumed to be matched if they have the same local point or area features. The matched visual appearances indicate that they are referring to the same object captured by different cameras.

So far, there are several point feature detectors and descriptors available, including SIFT [6,7] and SURF [8]. The point features from the SIFT are robust to the orientation changes of camera views, but it only detects the blob-like feature points, which might be sparse for the matching of object visual appearances in camera views. Compared with SIFT, SURF is detected faster, but they are not fully affine invariant. It means that little feature points could be found, when there is a significant change on the camera view orientations [11].

The area-feature based matching methods mainly rely on local image windows. Typically, the methods find seed points and propagate from these points into small image windows. Then, the matching could be conducted through the cross-correlation of the visual patterns in these windows. For example, Pratt [12] used the image intensities in the local windows as the patterns for the cross-correlation. Rashidi et al. [13] also used the adaptive color difference to match image windows. Compared with the point-feature based matching methods, the area-feature based matching methods could produce the dense matching results [14] and be robust to local affine distortions. However, the matching with the area features might still fail, especially when the local image windows did not contain distinctive visual patterns or the patterns contained were deformed due to the complex image transformations [15].

2.2. Spatial relationship-based matching

In addition to the reliance on the visual features, the relative spatial relationship of the objects of interest in camera views is also investigated to conduct the matching. One common spatial relationship for object matching is the epipolar geometry. According to the epipolar geometry, if the projection of a three-dimensional (3D) point X on the left view (X_L) is known, the corresponding epipolar line on the right view could be decided, and the projection of the point X on the right view (X_R) must be on the epipolar line, as shown in Fig. 1. Therefore, the search space for matching is restricted to a line [16].

Zhang et al. [17] relied on the Least Median of Squares (LMedS) to find the epipolar geometry between two camera views with an initial set of points. Based on the epipolar geometry, Lee et al. [4] proposed a method to match onsite construction workers captured in two camera views. Under their method, the location of each construction worker in the first camera view was used to determine its corresponding epipolar line in the second camera view. Then, the distances of the workers in the second camera view to the line were calculated. The one closest to the line was assumed to match the worker in the first camera view [4]. According to the tests conducted by Lee et al. [4] in real construction sites, the matching recall and precision could reach 71.4% and 98.7%. Recently, Konstantinou and Brilakis [18] combined the epipolar geometry under camera views with the shift of the workers' centroids and visual features across video frames to improve the matching accuracy.

2.3. Summary of existing matching methods in construction

Existing matching methods still have several limitations, when being used to match construction resources. This is partially because the video cameras are typically set up at height on construction sites. They shoot mobile equipment, workers, etc. on the sites far away. As a result, the size of these resources is small; and local visual features could not always be found to characterize their visual appearance in each camera view.

Also, existing matching methods based on visual features failed to match construction resources with similar visual appearances. For example, traffic cones are common on a construction site and they all look like each other. It becomes easy for the feature based matching methods to produce matching errors and fail to find the same traffic cones under different camera views.

As for the matching methods based on spatial relationship, they might fail, when there are two or more objects of interest along the same epipolar line in the camera view, as shown in Fig. 2. The failure was mainly due to the calculation of the epipolar lines. For example, the methods of Lee et al. [4] and Konstaninou and Brilakis [18] used the centroid of the bounding box of a worker to represent his/her location and calculate the corresponding epipolar line. The centroid of the bounding box does not always truly reflect the worker's location, especially when the worker in the first camera view is partially occluded. As a result, the deviations are introduced in the epipolar line calculation, which led to the matching error. Table 1 summarizes the