2D VISION TRACKING METHODS’ PERFORMANCE COMPARISON FOR 3D TRACKING OF CONSTRUCTION RESOURCES

Atefe Makhmalbaf1, Man-Woo Park2, Jun Yang3, Ioannis Brilakis4, Patricio A. Vela5

1 College of Architecture, Georgia Institute of Technology, Atlanta, GA. 30332-0355; PH (404)385-1276; email: atefe@gatech.edu
2 School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA. 30332-0355; PH (404)385-1276; email: mw.park@gatech.edu
3 Northwestern Polytechnical University, Xi’an, China, and School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250; PH (404) 894-3577; email: jyang67@mail.gatech.edu
4 School of Civil and Environment Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250; PH (404) 894-9881; email: brilakis@gatech.edu
5 School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250; PH (404) 894-8749; email: pvela@gatech.edu

ABSTRACT

In construction sites, tracking personnel, equipment, and materials is necessary in many applications such as asset management and progress monitoring. Vision tracking as a tracking technique has two unique capabilities which make it ideal for tracking on large scale, congested construction sites; tracking from a distance and tracking multiple elements from a single camera. 2D tracking results can be obtained through each camera and the 3D positions can be calculated by calibrating multiple cameras. 2D vision tracking methods vary and each method has unique capabilities, but little is known about the appropriateness of each method for tracking construction resources. This paper presents an evaluation of these methods that aims to identify the one most suitable for 3D tracking in construction sites. To satisfy this goal, 2D tracking methods are categorized, metrics are identified and performance of two categories of methods is evaluated based on these metrics. This evaluation has been used to find some preliminary results that can ultimately help with selection of the most suitable category of 2D vision tracking methods for construction applications.

KEYWORDS

Vision tracking, Image processing, Computer vision, Comparison
1 INTRODUCTION

In construction sites, tracking project related entities (e.g. personnel, equipment, and materials) is necessary in many construction applications such as progress monitoring, activity sequence analysis, productivity measurements, locating resources, and safety management. One potential capability of 3D vision tracking is that it can track tens of hundreds of objects that are present in the view of a camera without installation of any tags on them. 3D vision tracking is expected to be cost-effective and time-effective in large-scale, congested construction sites, where a large number of entities are required to be tracked. This is because of the low cost of cameras and also less pre-processing time needed to install cameras compared with other tracking methods that require installation of tags on each entity being tracked.

3D vision tracking results can be obtained by applying epipolar geometry that can calculate the depth information of objects using 2D vision tracking results (i.e. x,y information) from two or more cameras (Brilakis et al., 2008). The effectiveness of 3D vision tracking methods is directly related to the performance of 2D vision tracking since accurate positions (x, y) of objects being tracked are required for optimum 3D vision tracking results. Therefore, it is necessary to find a proper 2D vision tracking method out of the methods that currently exist to see which one can satisfy construction resource tracking needs and be adaptable to the construction site environment. Since little is known on how these methods compare for this environment, a scientific approach to select an appropriate 2D vision tracking method that can fit construction industry’s requirements can greatly enhance 3D tracking results.

There are a variety of 2D tracking methods with different capabilities and specifications in terms of the way of initializing the tracking (i.e. manual or automatic), the processing time, and their effectiveness in terms of handling occlusion, object size, etc. These aspects, along with complications that are in the nature of construction sites (e.g. existence of many different types of entities with high similarity in terms of shape and color (Brilakis et al., 2008)), make the selection of a proper 2D vision tracking method for construction applications a challenging task.

This paper presents a comparative study of existing tracking methods to find the most suitable one for construction site 3D tracking. To accomplish this objective, 2D tracking methods are categorized, studied and compared. A brief literature review shows that there are many different approaches and algorithms under each category of trackers, which results in enormous number of 2D tracking methods. This makes the literature so vast and no attempt can be nor will be made to review them comprehensively. Here, the authors have just tried to look at the main categories of trackers and tested some selected methods under each category to find the most optimum one for the purposes of 3D tracking in construction sites.

The comparison of 2D tracking methods is based on identification of criteria that can affect the efficiency of on-site tracking in construction sites which vary in terms
of their characteristics (e.g. size, complexity, congestion). These criteria have been set up as metrics (i.e. performance indicators) to evaluate and compare tracking methods for construction applications. Two categories of tracking methods (contour-based and kernel-based) were tested based on the number of frames each method can track. Preliminary results revealed that kernel-based tracker is capable of tracking more number of frames in each case.

2 VISION TRACKING

Object tracking has attracted a lot of interest and attention in the field of computer vision because of the large number of applications it has as well as availability of high performance computers and inexpensive digital video cameras. To achieve vision tracking, there are two main steps that must be accomplished. The first step is detection of objects of interest, which can be done manually or automatically depending on the method selected. This step has also been referred to as characterization or initialization step (Marfil et al., 2007). The second step is tracking of the detected objects and prediction of their changes in terms of position and shape in the succeeding image frame, which is also known as localization of targets (Marfil et al., 2007).

2D vision tracking is not sufficient for the general applications of tracking in construction industry and 4D information is required, i.e. 3D spatial coordinates over time. For instance, to avoid collisions at construction sites, the entities’ precise position (i.e. 3D) is needed in real time to calculate the relative distance and velocity of objects. In vision tracking, cameras can be used to obtain 2D positions of objects in each camera’s view. By integrating information collected from multiple cameras and applying epipolar geometry, it is possible to evaluate the depth information of objects (Brilakis et al., 2008). Consequently, obtaining reliable 3D tracking results depends on the results of 2D tracking. Therefore, selection of a 2D tracking method that is robust to domain specific variations of construction sites and produces valid tracking information is essential. 2D vision trackers can be classified as contour-based, kernel-based, and point-based trackers (Yilmaz et al., 2006). A brief review of each method is brought below.

2.1 Contour-Based Trackers

Many tracking methods use silhouette or contours to track objects by estimating and updating the region or boundary of the target in the current frame and comparing that with the results obtained from the previous frame (Nguyen et al., 2002). Contour-based tracking methods use a variety of algorithms; they may track parameterized contours and non-parameterized contours (Nguyen et al., 2002). Parameterized methods approximate the contour using a parametric model (e.g. B-Splines) or detect the contour by minimization of some energy function (i.e. snake model). Non-parameterized contours are basically represented as boundary of a
region (e.g. geodesic contour model) (Nguyen et al., 2002). The strength of the latter approach is that it can be used to represent the contour of an arbitrary shape (Nguyen et al., 2002) and as a result they have been used in most contour-based tracking methods of recent years.

Algorithms employed in contour-based trackers could use “photometric variables” (e.g. intensity, color, texture), “geometric variables” (e.g. edges, lines, corners), or a combination of both (Freedman 2004). It is argued that trackers that use photometric variables are more advantageous to geometric trackers because they are more reliable in the presence of illumination variations and cluttering and also they take into account the rich amount of information existing in the images (Freedman 2004). Based on literature, it could be safe to say methods that use non-parametric contours and photometric variables are more favorable to use because they are more robust to illumination variations as well as occlusion. The main advantage of contour-based trackers is that they can detect the exact contour, which is useful in applications where exact posture of the target is needed. Occlusion and illumination variations may not be handled as good as other categories of trackers.

One approach to achieve contour-based tracking is employment of variational frameworks (Yilmaz et al., 2006). These frameworks basically integrate variety of information for image segmentation Rousson and Deriche (2002). Bertalimo et al. (2000), Rousson and Deriche (2002), and Cremers (2003) presented different variational frameworks combining different information (e.g. motion estimation and shape regularization in Cremers (2003)). The framework proposed by Rousson and Deriche (2002) is a dynamical framework that works based on a Bayesian model. This method has shown very good results for image segmentation and contour evolution of different objects in a variety of backgrounds.

2.2 Kernel-based Trackers

Kernel basically refers to the shape and appearance of the object (Yilmaz et al., 2006). As a result, these trackers are also referred to as region-based or appearance-based methods. These methods compute and update the motion of the kernel in each frame based on the regions’ information such as color and texture (Marfil et al., 2007). Color histogram generally plays an important role in these methods since color is a good feature to distinguish objects’ regions and enables to manage partial occlusion (Marfil et al., 2007). However, the tracker may get confused if objects of similar colors are occluded by each other. Since these methods do not take any “a priori” knowledge of the objects’ appearance into account, initialization of tracking a new object has to be performed manually.

Two sub-categories of kernel-based trackers are template and density-based appearance models, and multiview appearance models (summarized in Table 1 below). Template-based models take into account both color histograms and spatial information. The basic concept of template tracking is finding the region that best
matches with the template that is determined in the first frame (Marfil et al., 2007). One approach to do this is mean-shift procedure, which basically is an analysis technique that works based on the comparison of histograms of the object, Q, with the window approximated around the location of that object, P (Yilmaz et al., 2006). Bhattacharya coefficient is used in this approach to measure the similarity of the two histograms. Templates and density-based trackers are advantageous because of their simplicity and low computational cost. However, they are not suitable if the pose of objects change considerably in each frame (Yilmaz et al., 2006).

2.3 Point-based Trackers
In point-based trackers, objects are represented by some limited points, which are detected and tracked in consecutive frames (Yilmaz et al., 2006). In image frames that objects appear larger, the more points may be demanded for successful identification of the object, which could affect processing time and also when multiple points are employed, grouping of points that belong to the same object becomes an important problem (Yilmaz et al., 2006). Also, it is important to select and keep appropriate feature points that can effectively represent the objects. Yilmaz et al (2006) have divided point-based methods into two categories of deterministic methods and probabilistic methods. In general, point-based trackers using multiple points are beneficial to characterization of non-rigid objects provided that the locations of the feature points are flexible. Also, these trackers are suitable for tracking extremely small objects which can be represented by a single point, but they have more tendencies to fail in the presence of occlusion as well appearance and re-appearance of objects (Yilmaz et al., 2006).

2.4 Summary of Methods
The following table presents a summary of tracking categories outlined above. This table is based on the literature review and especially the survey of object tracking methods done by Yilmaz et al. (2006). It should be noted that the general strengths and limitations listed in this table are addressed and handled by some subcategories/approaches and it cannot be concluded that a category has such limitation regardless of the algorithm or method used. For example, although tracking multiple objects cannot be handled by most kernel-based trackers, there are some algorithms that can track multiple objects.
Table 1: A summary of Main Features of Object Tracking Methods

<table>
<thead>
<tr>
<th>Category</th>
<th>Main Purpose</th>
<th>General Strength(s)</th>
<th>General Weakness(es)</th>
<th>Performance Evaluation</th>
<th>Sub-categories</th>
<th>Approaches (examples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contour-based</td>
<td>Tracking exact contour (object shape)</td>
<td>Ability to represent variety of shapes</td>
<td>handling occlusion and object split and merge</td>
<td>precision and recall</td>
<td>Contour-evolution</td>
<td>Variational Methods</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Matching Shapes</td>
<td>Histogram</td>
</tr>
<tr>
<td>Kernel-based</td>
<td>estimating object motion (region and direction)</td>
<td>handling occlusion</td>
<td>handling multiple objects</td>
<td>precision and recall</td>
<td>Template-based models</td>
<td>Mean-shift</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>distance b/w estimated and actual</td>
<td>Multi-view models</td>
</tr>
<tr>
<td>Point-based</td>
<td>tracking small objects (single point tracking)</td>
<td>tracking small objects</td>
<td>Handling occlusion and appearance &amp; re-appearance</td>
<td>Precision and recall</td>
<td>Deterministic</td>
<td>MGE Tracker</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Statistical</td>
<td>MHT filter</td>
</tr>
</tbody>
</table>

Fig. 1: Representation of different categories of trackers. a) contour-based b) kernel-based and c) point-based trackers from Yilmaz et al., (2006).

3 COMPARISON MATRICS

Construction sites have some unique characteristics, which should be considered when it comes to object tracking, because they affect the appearance of objects in the view of the cameras. The scale of construction sites, the congestion, the clutter, the variety of objects with high similarity in terms of color and shape, the illumination as well as the illumination variations and the existence of both rigid and non-rigid objects (e.g. equipment and workers) that enter and leave the site are the main characteristics of most construction sites. Therefore, to find the most suitable 2D tracking method, one needs to know how well each of these factors can be handled. To satisfy this, a list of metrics that are based on these characteristics is proposed here to evaluate and compare the 2D vision tracking methods presented previously. These metrics are based on the ability of the tracker to handle: 1) occlusions, 2) small objects, 3) objects of different types, 4) illumination variations; 5) absolute value of illumination, and also 6) processing time.

Occlusion refers to the state when an object is partially or fully blocked by itself or another object. In construction sites, there is constant movement of equipment,
materials and personnel. As a result, it is extremely important to select a method that can continue to track objects even in the presence of occlusions. Self-occlusion is a common problem of contour-based methods especially when tracking non-rigid objects. It should be noted that one way to handle and control occlusions is by picking the right position to install cameras. By installing the cameras in higher levels, it is possible to minimize complete occlusion of objects; partial occlusions can be handled well by some methods (e.g. kernel-based).

Small objects can be defined as distant objects that appear to be small in the view of the camera and also those that have smaller dimensions compared to other objects (e.g. hardhats vs. bulldozers). Construction sites vary in size and some objects appear to be very small in the view of camera. Also, some objects move away from the camera and become small in size. Therefore, it is important to select a method that can track small objects and also be able to handle changes in scale.

The types of objects in construction sites refer to their rigidity and non-rigidity of objects. Rigid objects (e.g. wheel loaders, backhoes, steel beams) do not make severe changes in their poses. Therefore, their appearances are adequately predictable. On the other, non-rigid objects (e.g. workers) change their poses dynamically. This makes their appearances in the subsequent frames unpredictable. In most construction related applications, there is no need to track the whole body of a worker. As a result, it is possible to overcome the difficulty of tracking non-rigid objects by tracking a rigid part of their belongings (e.g. hard hat).

Illumination variations are related to changes that occur in lighting conditions. Such variations can affect tracking results because they can significantly alter the appearance of objects in the view of the cameras. Illumination variations are common at open, large-scale and congested construction sites, where shadows and light reflections (from other equipment or adjacent buildings) are present. Therefore, the ability to handle illumination variations is an important factor to be considered.

The absolute value of illumination is another factor related to lighting conditions. Absence of light or presence of strong illumination can affect tracking results depending on how severe each case is. Working hours at construction sites can start from very early hours in the morning when it is still dark and go on until very late hours of the day. Also, construction site locations vary geographically from very illuminated sites at deserts to light reflecting sites by oceans. Therefore, it is important to select a tracker that can perform accurately at different light levels in different locations and during the course of a day.

Processing time can be defined as the amount of time it takes a tracker to process one image frame and send the results to the user. After receiving and processing video data, the tracker should be able to transfer the accurate information to the user in a specified amount of time. For instance, when tracking for safety management, after the tracker detects a short distance that could result in a collision between a bulldozer and a worker, it should warn the worker before the collision occurs. Different
applications have different requirements for 2D vision tracking methods’ processing time and the tracker that has less response time than the requirement should be selected.

4 RESULTS AND COMPARISON

The performance of tracking categories was tested and evaluated based on the metrics discussed in the previous section. For comparison purposes, a kernel-based method was selected that uses the Bhattacharyya coefficient as similarity measure and the mean-shift algorithm to carry out the optimization based on a method presented by Comaniciu et al., 2003. Also, a variational framework was selected as a contour-based tracker. Rousan and Deriche (2002) used a Bayesian model to develop their variational method and the authors combined this framework with a knowledge-based segmentation algorithm presented by Haker et al. (2001). These methods were implemented and tests were performed using videos that were taken at construction sites.

Some frames from preliminary tracking test results are shown in Fig. 2 below. In the first row of images, the worker (non-rigid object) is being tracked using the contour-based method. This worker could be tracked for 10 frames until he was occluded by another worker and then the tracking failed. On the other hand, the result from the kernel-based method shows that this method could still track the worker in the equivalent frame and even the subsequent frames. In the second row of images in Fig. 2, a bulldozer that is moving away from the camera is being tracked. The results show that the contour-based method loses the target in frame 2500 while the kernel-based tracker keeps tracking for another 6000 frames (until the object comes to the camera). The last row of images shows the results of tracking of a backhoe’s bucket. Complete deformation of the bucket in the view of the camera can be observed in these images. The results revealed that the kernel-based method could track the bucket for 723 frames while the contour-based lost it in frame 101. It should be noted that illumination varies in each frame and affects the appearance of the target in the view of the camera. Therefore, both methods can handle illumination variations to a good degree. The difference in the processing time of these methods was not that significant and as a result it is considered to be a neutral criterion for the purposes of this evaluation and comparison.

To evaluate the performance of tracking methods, it is possible to measure the accuracy of each method by calculating their precision and recall as mentioned in Table 1 in section 2.4 above. It is also possible to assess the performance of tracking methods based on the number of frames successfully tracked to obtain some preliminary results. The authors have closely watched the frames and recorded the number of frames the object can be tracked using each tracking method. A summary of this comparison is shown in Table 2 for each case and criterion.
5 CONCLUSION

Tracking objects in the construction sites is essential in many applications such as progress monitoring, activity sequence analysis, productivity measurements, locating resources, and safety management. These applications require the results of 3D vision tracking over time while calculating the depth information of objects (the 3rd dimension) depends on the results of 2D vision tracking. Selecting the best 2D vision tracking method for construction applications faces some challenges and difficulties. First of all, there are a number of 2D vision tracking methods with different capabilities and specifications. In addition, construction sites have unique tracking requirements, which must be considered in order to select a 2D tracking method that performs optimal under restrictions of construction sites.

This paper presented a systematic approach to compare different categories of 2D vision tracking methods in construction environments. The comparison was conducted using a set of criteria such as occlusion handling, small object detection, object type variability, and illumination variations. The results showed that the Kernel-based (Mean-shift) method performed better than the Contour-based (Variation method) in most cases.

Table 2. Number of frames successfully tracked.

<table>
<thead>
<tr>
<th>Object Type</th>
<th>Kernel-based (Mean-shift)</th>
<th>Contour-based (Variation method)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occlusion (e.g. worker shown in images a-d)</td>
<td>115</td>
<td>102</td>
</tr>
<tr>
<td>Small object and also scale variation (e.g. road roller shown in images e-h)</td>
<td>8564</td>
<td>2500</td>
</tr>
<tr>
<td>Object type and quick deformation (e.g. bucket of a backhoe shown in images i-l)</td>
<td>723</td>
<td>101</td>
</tr>
<tr>
<td>Illumination variation (e.g. bucket of a backhoe shown in images i-l)</td>
<td>723</td>
<td>101</td>
</tr>
</tbody>
</table>
vision tracking methods to find the most optimum one for 3D vision tracking purposes. The 2D vision tracking methods were first categorized and each was briefly described. The characteristics of construction sites that can interfere with the performance of these vision tracking methods were recognized and comparison metrics were set up based on them, to measure the performance of each category in relation with construction site requirements. One method from contour-based and another from kernel-based trackers was selected and implemented. A number of tests under different conditions of construction sites were performed to observe the performance of each tracking algorithm. The methods were compared based on the number of frames each tracker could successfully track. The preliminary results of this comparison revealed that kernel-based methods are more reliable between the two categories of trackers tested for construction related applications.

Future work will focus on measurement of accuracy of each method under each condition. Point-based methods will also be implemented, tested and compared with the results obtained in this study. Ultimately, the preferred 2D tracking method will be used to implement 3D vision tracking systems by employing epipolar geometry algorithm.

ACKNOWLEDGEMENT

The authors would like to thank the generous support of National Science Foundation through projects 0933931 and 0904109. Also, they wish to gratefully acknowledge the contributions and help of Ibrahima J. Ndiour and Omar Arif from School of Electrical and Computer Engineering, Georgia Institute of Technology.

REFERENCES


