INITIALIZING VISION BASED TRACKERS USING SEMANTIC TEXTON FORESTS

Man-Woo Park*, Gauri M. Jog, and Ioannis Brilakis

Department of Civil & Environmental Engineering, Georgia Institute of Technology, Atlanta, USA
** Corresponding author (mw.park@gatech.edu)*

ABSTRACT: Vision based tracking can provide the spatial location of project related entities such as equipment, workers, and materials in a large-scale congested construction site. It tracks entities in a video stream by inferring their motion. To initiate the process, it is required to determine the pixel areas of the entities to be tracked in the following consecutive video frames. For the purpose of fully automating the process, this paper presents an automated way of initializing trackers using Semantic Texton Forests (STFs) method. STFs method performs simultaneously the segmentation of the image and the classification of the segments based on the low-level semantic information and the context information. In this paper, STFs method is tested in the case of wheel loaders recognition. In the experiments, wheel loaders are further divided into several parts such as wheels and body parts to help learn the context information. The results show 79% accuracy of recognizing the pixel areas of the wheel loader. These results signify that STFs method has the potential to automate the initialization process of vision based tracking.

Keywords: *Artificial intelligence, Automatic identification, Image processing, Tracking, Information technology*

1. INTRODUCTION

In recent years, state-of-the-art information technologies (IT) have been introduced to construction engineering in order to automate data collection on construction sites. These technologies attempt to provide real time information which enables faster and more informed decisions in construction management. One of the active research areas related to automated data collection in construction sites is tracking of construction resources, which generates real time information of the resources' positions. Various tracking methods which include Radio Frequency Identification (RFID), Global Positioning Systems (GPS) and Ultra Wideband (UWB) have been tested and applied to construction entities [1, 2, 3]. However, the large number of entities to track on large scale, congested construction sites limits the applicability of these technologies. Since they require installing tags or sensors on the entities to track, they are expensive in terms of time and cost. The installation process also creates privacy issues when tracking workers who do not want to be tagged.

Vision-based tracking proposed by Brilakis et al. [4] has a great potential to track construction entities in large scale, congested sites since it can track a large number of entities with only cameras. As shown in Fig. 1, this method uses two camera views in which construction entities are tracked with general 2D vision tracking algorithms. The relations between two views are discovered by camera calibration, and the 2D tracking results from the two views are correlated by entity matching and triangulation processes, which finally compute 3D positions. In order to initiate 2D tracking, the pixel regions of entities to track have to be determined in the first frame of the camera views. Allowing for a large number of construction entities on site, a method to automatically detect the entities is required.

This paper presents the application of an object detection algorithm for the initialization of 2D tracking. Various kinds of object detection algorithms are investigated, and STFs method [5] is chosen for detecting wheel loader which is one of the typical construction equipment. In the experiment, 35 images are used for training, validation, and testing. By appropriately dividing the wheel loaders into several detailed parts, 79% of pixel-wise accuracy is achieved, which signifies that it is promising to use STFs to automate the initialization of vision based tracking.

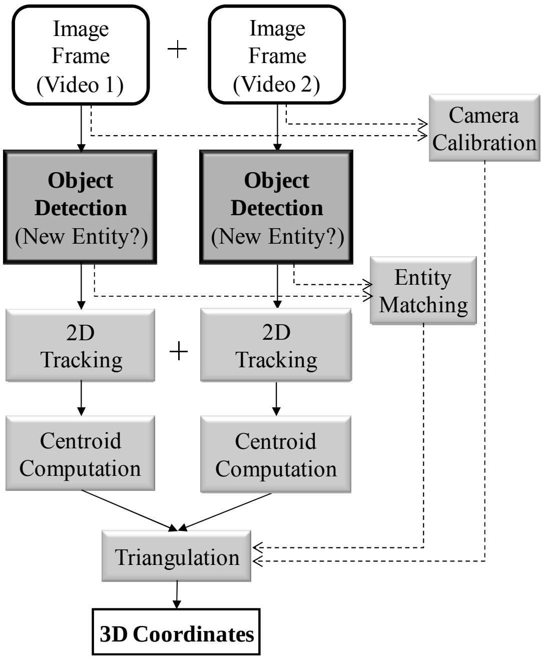

Fig. 1 The flow chart of 3D vision tracking.

2. BACKGROUND

Object detection is a typical research topic in computer vision area. Various kinds of approaches have been proposed and applied to diverse objects. When it comes to the initialization of tracking, the objects to detect are in motion. Background subtraction [6, 7] is a famous object detection algorithm that takes object motion into account. It constructs a static background model, and compares incoming frames with the model. The moving objects are detected by masking the pixel areas that is far different from the background model. The capability to detect all moving objects with a single process makes background computationally efficient. However, it cannot sort out the entities which are not related to the construction project (e.g. pedestrians, birds, cars). Also, when multiple objects appear partially overlapped or too close to each other, they can be detected as one object. The sensitivity to the changes in illumination conditions also limits its applications.

Template matching constructs an appearance-based model to represent an object type, and compare it with models obtained from window areas of the test image. The comparison is performed using a convolution mask. The convolution results will give highest values at places where the models best matches. The template model can be composed of point features [8], edge features [9], or kernel features [10]. The limitation of using point features is difficulty in grouping points that belong to the same object. When employing edge features, it is important to differentiate edges of the objects' shadows which appear variously depending on illumination conditions. Appearance-based template matching methods that use kernel features such as wavelet, Gabor filters, and Haar-like features have been successfully applied to vehicle detection [11, 12, 13]. It is expected to work well for construction equipment since it has similar appearance to vehicle in terms of rigidity and angular characteristics (straight lines of boundaries). Sun et al.'s method [12] took advantage of both wavelet and Gabor features for vehicle detection. It extracted straight lines using wavelet features, and made the extracted line features adaptable to both orientation and scale variation with Gabor features. Their experiment results showed accuracy enhancement.

Since most objects are viewed differently depending on in viewpoint, perspective, illumination conditions, and occlusions, it is not effective to represent various appearances with one template model. Also, in construction sites, there are various types of resources to track (e.g. workers, backhoes, loaders, precast beams). To deal with all types, template matching needs as many matching process as the number of object types. To overcome this problem, machine learning process is required. In this process, a number of template models extracted from training images are learned. It trains a classifier with the training images include both positive images (images that include the object to detect) and negative images (images that are not related to the object).

Recently simultaneous process of segmentation and recognition has become prevalent for object recognition

[14, 15]. It employs machine learning with feature descriptors such as SIFT (Scale Invariant Feature Transform) [16] and SURF (Speed-Up Robust Features) [17] from local patches. The collections of descriptors are then trained to create the codebook of visual words. Based on the codebook, a word is assigned to each image patches. They segment images and categorize each segment into pre-defined categories. Shotton et al. [5] proposed an algorithm known as the STFs method that is based on a kind of kernel-features instead of feature points.

In this paper, the STFs method is employed for the detection of wheel loader. The STFs method uses bag of textons in which textons were obtained by the raw pixel values or sum, difference of the pair of pixels in the specified box surrounding each pixel. Also, it trains context information of the objects through the region prior distribution. These characteristics of the STFs method are taken into account and judged to be effective to recognize wheel loaders.

STFs are randomized decision forests trained with the bag of semantic textons. The semantic texton is computationally less expensive than SIFT or SURF since it uses basic raw pixel values. Semantic texton is a vector of pixel values, the sum and the difference of the two far-off pixel values in a fixed size of rectangular window around a pixel. It does not require detecting corner points or calculating computationally complicated feature descriptors (e.g. histogram of gradients). However, the use of pixel values in the rectangle patch itself is not invariant with rotation and scale [5]. Therefore, it can be resolved by training scaled and rotated images. The original training images are replicated with scaling and rotating. The replicated images are added to the training images. Randomized decision forest is a classifier that consists of multiple binary decision trees [18]. Each node provides the probability of a class given that the process reaches the node. Also, it decides whether to go down to the left or right child node based on the raw pixel values.

3. METHODOLOGY

STFs method is chosen considering the general appearance of wheel loaders. The Figure 2 shows general appearance of wheel loaders. Wheel loader can be divided into several parts that have distinct characteristics. Wheels have circular shapes and also specific texture patterns. Top parts (where operators sit) have yellow or black surfaces, and body parts (front and rear parts) are mostly yellow. Buckets have unique shapes, but the appearance of the shape varies depending on the perspective view. Also, depending on the equipment, various colors of bucket can be seen. Black and yellow are basic colors of the bucket. However, while getting worn, rusty and dirty, the appearance of bucket changes. In terms of arrangement of parts, wheels usually appear on the bottom part of the wheel loader and body part above the wheels, the top part above the body part, and buckets next to the body part.

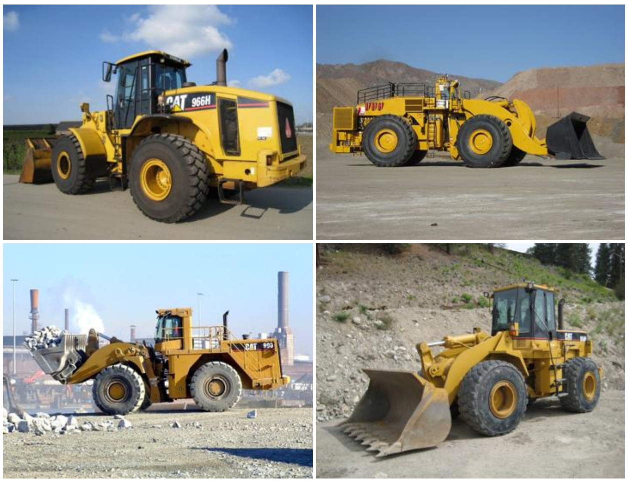

Fig. 2 Appearances of wheel loaders.

In order to train the forest, labeled images are required. Examples of the labeled images are shown in Figure 3, in which different colors indicate different labels. A label is assigned to each pixel. The decision forests are created through training of each pixel's semantic texton feature with the assigned label. In this paper, three types of labeling strategies are tested. First, only 2 classes are included in the labeling – whole region of a wheel loader as 'wheel loader' and the remaining region as 'background' (Fig. 3(b)). Second, 4 classes are included in the labeling – 'wheel', 'body', 'top', and 'background' (Fig. 3(c)). Third, 6 classes are included in the labeling – 'wheel', 'front body', 'rear body', 'top', 'bucket', and 'background' (Fig. 3(d)). The second and third strategies are made based on the assumption that each part has specific distinct features that are good to distinguish from others. Also, since STFs method trains context information as well as semantic texton feature itself, the detailed labeling can add more information about the relative positions of the divided parts. In the second case, the front and rear body parts which are treated separately in the third case are categorized as one single part since they are more similar to each other than to other part. The bucket is removed from the wheel loader based on the inference that appearances of the bucket are so diverse that inadequate training may harm the detection performance. Furthermore, the area of the other parts than the bucket is large enough to track wheel loaders with 2D vision tracking.

As mentioned in the previous section, STFs method generates semantic texton feature in fixed size of windows. If training images have far different sizes of wheel loader, the training may not be able to form common characteristics of wheel loader despite additional replicated images with scaling. Hence, all images are resized to have similar sizes of wheel loader, which can facilitate the training. In addition, it is necessary to find an optimal window size that allows extracting better features of wheel loader given the image sizes of wheel loader. Three values of window sizes are tested and compared in our experiment.

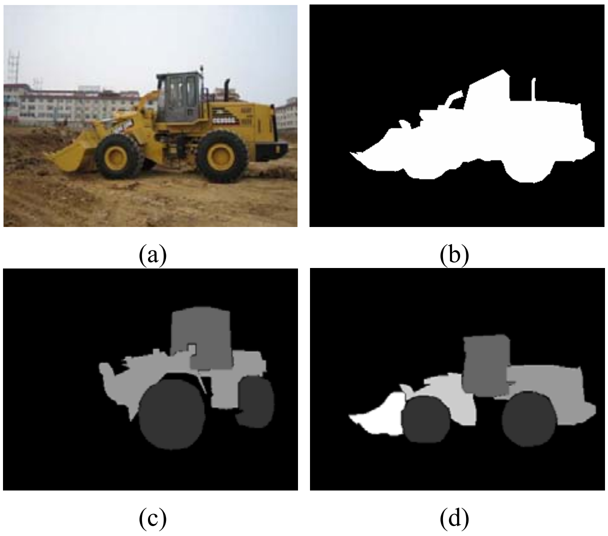

Fig. 3 The example of labeling: (a) a training image (b) 2 classes (c) 4 classes (d) 6 classes labeling.

4. EXPERIMENT RESULTS

In the experiment, the STFs method is applied to 35 images that include wheel loaders. Various objects such as soil, sky, trees, building, rocks and woods are in the background of the images. Out of 35 images, 17 images are used for training and 2 images are used for validation. The remaining 16 images are used for testing which gives segmentation results and pixel-wise accuracy.

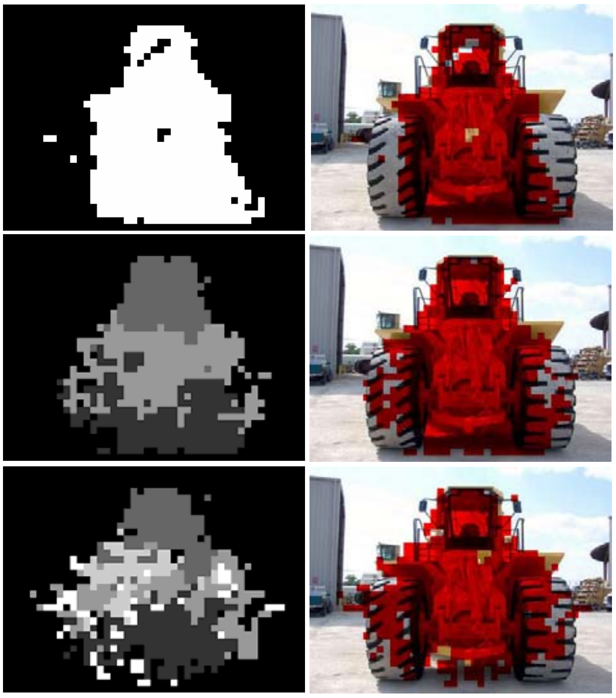

Fig. 4 Results of loader's rear view detection.

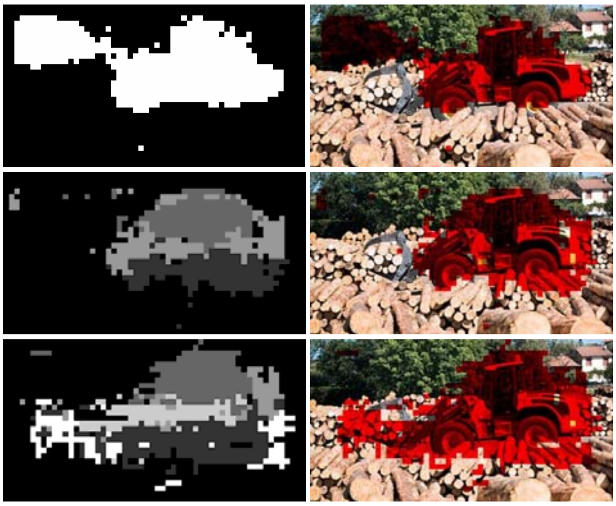

Fig. 5 Results of loader's side view detection.

Figure 4 and 5 show examples of the results. The first column is presenting the segmentation and categorization results and the second column is showing the detection results (darken area) based on the first column's result. In Figure 4, the results of 4 classes and 6 classes labeling shows more accurate segmentation than the 2 classes labeling. 2 classes labeling fails to detect wheels. The difference between the features of wheels and other parts causes this failure.

In Figure 5, 2 classes labeling completely misses the bucket, and detects bucket-shaped, dark region of the trees. When 4 classes labeling is used, it recognizes the wheel loader very well without detecting the bucket (in four classes labeling bucket is not trained). When 6 classes labeling is used, it recognizes the area in front of the body as a bucket even though the detected part is actually timbers. This proves the fact that STFs method learns the context information. In this case, it is inferred to learn the context information that bucket usually appears in front of the body. In segmentation results of Figure 4 and 5, it can be seen that wheel is detected at the bottom part which also indicates the learning of context information.

Table 1 shows the pixel-wise accuracy of three labeling strategies. The pixel-wise accuracy is calculated by dividing the number of correctly segmented pixels by the size of the images. STFs method gives the best results of 79.0% accuracy with 4 classes labeling.

Table. 1 Pixel-wise accuracy for three kinds of labeling

Labeling strategies	2 classes	4 classes	6 classes
Pixel-wise accuracy	70.6%	79.0%	71.5%

As described in previous section, three different values of window size - 5 x 5, 10 x 10, and 15 x 15 - are tested to find optimal one, given the wheel loader size in the images (ranging approximately from 200 x 100 to 300 x 200). Figure 6 shows an example of the test with 4 classes labeling, and Table 2 presents the pixel-wise accuracies. In Figure 6, the result of using 10 x 10 window size covers more regions of wheel loader than other results. Also, it gives higher accuracy than 5 x 5 or 15 x 15.

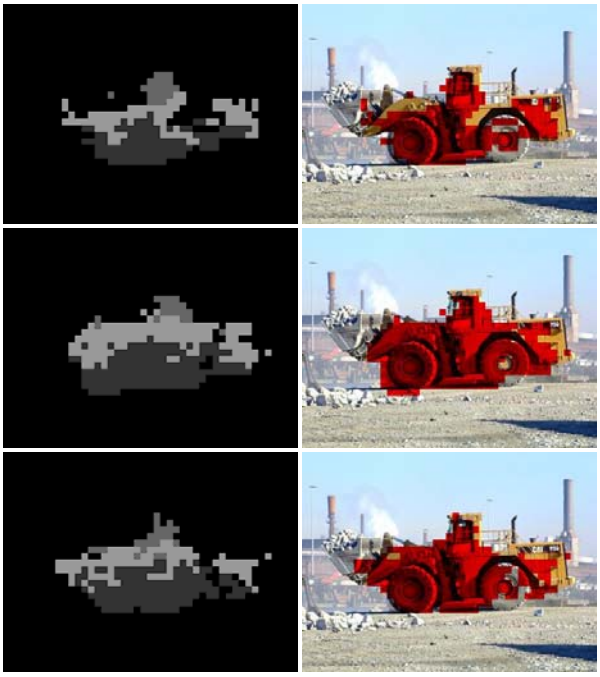

Fig. 6 Results of loader's side view detection.

Table. 2 Pixel-wise accuracy for three kinds of window sizes

Window size	5 x 5	10 x 10	15 x 15
Pixel-wise accuracy	77.8%	79.0%	78.6%

5. CONCLUSIONS

In vision tracking methods' framework, object recognition methods are required in order to automate the initialization process which locates the objects at first frame. STFs method is employed for wheel loader detection. STFs method mainly uses bags of textons and randomized decision forests and it learns context information as well as the localized appearance features. Three cases of labeling strategies which reflect wheel loader appearances with different extent of details are tested and compared. Also, three values of window sizes are tested to figure out the effect of this parameter on the results and find the optimal one. 4 classes labeling that divides wheel loader regions into top, body, and wheel, is found the best for extracting distinct features of wheel loader's parts. 10 x 10 window size gives slightly better result than 5 x 5 or 15 x 15. Even though the pixel-wise accuracy is a little less than 80 %, the detected regions are large enough to initialize wheel loader area for 2D vision tracking.

As a future work, a large scale of dataset needs to be created for various type of equipment. Then, simultaneous detection of different types of construction equipment will be tested.

REFERENCES

[1] Goodrum, P. M., McLaren, M. A. and Durfee, A., "The application of radio frequency identification technology for tool tracking on construction job sites", *Automation in Construction*, Vol. 15(3), pp. 292–302, 2006.

[2] Ergen, E., Akinci, B. and Sacks, R., "Tracking and locating components in a precast storage yard utilizing radio frequency identification technology and GPS", *Automation in Construction*, Vol. 16, pp. 354-367, 2007.

[3] Saidi, K. S., Teizer, J., Franaszek, M., and Lytle, A. M., "Static and dynamic performance evaluation of a commercially available ultra wideband tracking system", *Automation in Construction*, 2010.

[4] Brilakis, I., Park, M.W. and Jog, G., "Automated Vision Tracking of Project Related Entities", *Journal of Advanced Engineering Informatics*, Elsevier, in press, 2011.

[5] Shotton J., Johnson M., and Cipolla R., "Semantic Texton Forests for Image Categorization and Segmentation", In Proc. *Int. Conf. Computer Vision and Pattern Recognition*, June 2008.

[6] Ervin, R., MacAdam, C., Walker, J., Bogard, S., Hagan, M., Vayda, A., and Anderson, E., "System for assessment of the vehicle motion environment (SAVME)", University of Michigan Transportation Research Institute, Ann Arbor, Michigan, 2000.

[7] Melo, J. and Naftel, A., "Detection and classification of highway lanes using vehicle motion trajectories", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 7(2), pp. 188-200, 2006.

[8] Kim, Z. and Cao, M., "Evaluation of feature-based vehicle trajectory extraction algorithm", *13th International IEEE Annual conference on Intelligent Transportation Systems*, pp. 99-104, 2010.

[9] Kim, Z. and Malik, J., "Fast vehicle detection with probabilistic feature grouping and its application to vehicle tracking", In Proc. *the 9th IEEE International Conference on Computer Vision*, pp. 524-531, Nice, France, 2003.

[10] Troung, Q. B. and Lee, B. R., "Vehicle detection algorithm using hypothesis generation and verification", In Proc. *the 5th International Conference on Emerging Intelligent Computing Technology and Applications*, pp. 534-543, 2009.

[11] Schneiderman, D., "A statistical approach to 3D object detection applied to faces and cars", CMU-RI-TR-00-06, *Robotics Institute*, Carnegie Mellon University, May 2000.

[12] Sun, Z., Bebis, G. and Miller, R., "Improving the performance of on-road vehicle detection by combining gabor and wavelet features", *The IEEE 5th International Conference on Intelligent Transportation Systems*, pp. 130-135, Singapore, 2002.

[13] Haselhoff, A., Schauland, S. and Kummert, A., "A signal theoretic approach to measure the influence of image resolution for appearance based vehicle detection", In Proc. *the IEEE Intelligent Vehicles Symposium*, pp. 822–827, June 2008.

[14] Uijlings, J.R.R., Smeulders, A.W.M., and Scha, R.J.H., "Real-Time Visual Concept Classification", *IEEE Transactions on Multimedia*, Vol. 12(7), pp. 665-681, 2010.

[15] Zhang J., Marszałek M., Lazebnik S., and Schmid C., "Local features and kernels for classification of texture and object categories: A comprehensive study", *International Journal of Computer Vision*, Vol. 73(2), pp. 213–238, 2007.

[16] Lowe, D. G., "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision*, Vol. 60(2), pp. 91-110, 2004.

[17] Bay H., Tuytelaars T., and Gool L.V., "Surf: Speeded up robust features", *In Computer Vision - ECCV*, (A. Leonardis, H. Bischof, and A. Pinz, eds.), Vol. 2951, pp. 404-417, 2006.

[18] Geurts, P., Ernst, D. and Wehenkel, L., "Extremely randomized trees", *Machine Learning*, Vol. 36(1), pp. 3–42, 2006.