MOTION-BASED BACKGROUND SUBTRACTION AND PANORAMIC MOSAICING FOR FREIGHT TRAIN ANALYSIS

Avinash Kumar, John M. Hart, Narendra Ahuja

University of Illinois at Urbana-Champaign,
Beckman Institute for Advanced Science and Technology,
{avinash,jmhart3,n-ahuja}@illinois.edu

ABSTRACT

We propose a new motion-based background removal technique which along with panoramic mosaicing forms the core of a vision system we have developed for analyzing the loading efficiency of intermodal freight trains. This analysis is critical for estimating the aerodynamic drag caused by air gaps present between loads in freight trains. The novelty of our background removal technique lies in using conventional motion estimates to design a cost function which can handle challenging texture-less background regions, e.g. clear blue sky. Supplemented with domain knowledge, we have built a system which has outperformed some recent background removal methods applied to our problem. We also build an orthographic mosaic of the freight train allowing identification of load types and gap lengths between them. The complete system has been installed near Sibley, Missouri, US and processes about 20-30 (5-10 GB/train video data depending on train length) trains per day with high accuracy.

***Index Terms*—** background removal, panoramic mosaicing, intermodal freight train, wayside inspection

1. INTRODUCTION

Intermodal (IM) freight trains are the most common and economical mode of transporting goods across long distances in the North American Freight Railroads network. These trains are composed of different kinds of loads mounted and placed securely on rail cars. They operate at speeds of $75-80$ miles per hour (mph). At such high speeds, the air drag between the gaps of loads creates considerable amount of air resistance resulting in increased fuel consumption and operating costs. It was shown in [1] that an analysis of loading efficiency and gaps in an IM freight train can help railroad companies evaluate their loading techniques at IM facilities and save fuel costs.

A machine vision system which can compute length of all gaps present in an IM freight train by analyzing a video of the train was proposed in [2, 3]. The background (BG) in such videos consists of trees, sky etc visible through the gaps and above the train (Fig. 1(a)). The foreground (FG) consists of the fast moving train (Fig. 1(b)). An accurate BG subtraction in these videos allows us to identify gaps and boundaries of loads on the train. In the current setting of IM train videos, the problem of BG subtraction is made challenging due to the following constraints and requirements:

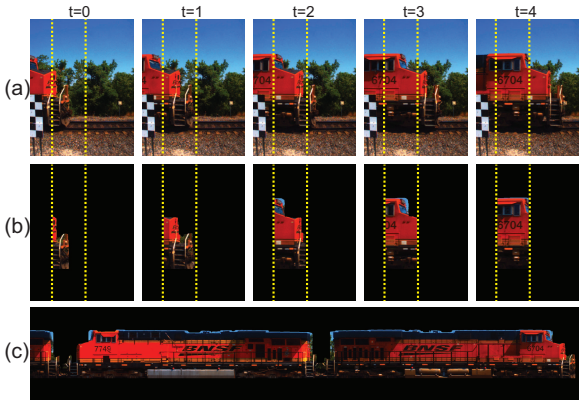

Fig. 1. (a) Consecutive image frames of an intermodal train video with moving trees visible in the background. (b) Background removed using our technique. (c) Orthographic mosaic of the train.

Accuracy: FG needs to be detected accurately for correct gap estimation between successive loads.

Image distortion: Radial distortion and perspective projection cause similar scene points to image differently in consecutive frames of the video. This change in shape of FG objects is more prevalent when the object is close to the camera, e.g. loads in IM train videos.

Illumination variations: Long term and short term changes in weather conditions and sunlight direction can modify the captured intensity of FG and BG in a single video.

Camouflage: The FG and the BG can be of similar color making it difficult to distinguish them.

Computational Speed: Since the goal is to develop a computer vision application for real world use, the computational speed of BG subtraction is critical. Typically 30-40 trains need to be captured and processed per day, where each train is captured at 30 fps and requires approximately $5-10$ GB of

Thanks to BNSF Railways.

storage. Thus the BG removal has to be fast while also being accurate.

Image noise: Photon noise and sensor noise are inherent in acquired videos as the aperture is kept open for short periods to avoid motion blur due to fast moving trains.

In this paper, we focus on developing a BG subtraction method which can handle aforementioned issues. Traditionally intensity modeling [4, 5, 6] have been used for BG subtraction. We employ a motion estimation based technique for BG subtraction as they can handle persistent dynamic nature of BG and FG [7] e.g. train in our videos. But motion estimation is known to fail at texture-less regions [8] e.g BG consisting of clear sky. Thus, we need to define a new cost function which can handle such situations. In addition, we also show a technique to create panoramic mosaic of the complete BG removed train. Our contributions are:

1. Designing a motion based cost function (Sec. 3) which is robust than naive motion estimates to distinguish static BG from dynamic FG. Specifically, it can handle texture-less regions which are known to be challenging for motion estimation [8]. This simple method has just a single parameter τ (Sec. 3.2) which needs to be set manually as compared to other sophisticated methods which typically have multiple parameters. The upside of this is that we can handle many videos with varying illumination conditions while still obtaining accurate BG subtraction results (Sec. 5).

2. Generating an orthographic panoramic mosaic (Sec. 4) of the train using motion estimates and BG removed images. This is useful for gap detection, visualization and classification of loads on the train [2].

2. PREVIOUS WORK

Background subtraction is a popular and well studied problem in computer vision. We present the prior work with respect to generic BG subtraction and domain dependent BG removal pertaining to IM freight train analysis.

Generic BG subtraction: Many techniques have been developed for generic BG subtraction [9, 10]. The most common technique is to model pixel intensities as a time series and fit a dynamic unimodal or multi modal Gaussian distribution to them [11, 4, 12]. Elgammal [13] proposed a non-parametric modeling of BG distribution based on kernel-density estimation. All of these techniques appear to fail and become parameter sensitive if the FG and BG are similar in intensity. This affects the applicability of these techniques on videos captured under wide range of illumination conditions. The BG subtraction problem can also be modeled as FG extraction by employing motion based features to distinguish fast moving FG and static/quasi-static BG [7, 14].

Intermodal BG subtraction: For IM freight train analysis, Kumar [2] did BG subtraction using simple edge detection techniques. But this technique required appropriate values for a number of parameters making it unsuitable for handling wide range of videos. This was followed by a statistical learning based approach in [3], which employed domain knowledge to learn background removal parameters but still the sensitivity of this algorithm to its parameters made it difficult to generalize to different background and illumination conditions throughout an year.

3. MOTION BASED BACKGROUND SUBTRACTION

In this section, we describe a hypothesis and validation based technique for BG removal using motion based features for IM freight train videos. We define this feature at each pixel location in an image frame as the amount by which this pixel shifts horizontally across consecutive frames. The vertical motion is assumed to be negligible. This is imposed by calibrating the pose of the camera such that there is only horizontal motion of the train. We also assume that there is at least a single frame of complete BG visible before the train appears in the video. This provides us with a model for the BG.

A video with N image frames is denoted as V. Thus $V = \{I_1, \ldots, I_t, \ldots, I_N\}$ where I_t is the image frame at time index t. Let us consider I_t from which we want to remove BG. The hypothesis and validation steps are as follows.

Hypothesis: As the camera is static and the rails of the track are fixed, we know the location of the moving railcar and wheels of the train in a image of the train. Thus, we know the location of some regions of the FG. We select two image patches A and B at the known FG location in I_t and the next image frame I_{t+1} respectively. The height of patch A and patch B are same but patch B is wider than patch A. The patch A is then correlated with shifted versions of patch B and the shift which results in maximum correlation is computed. This shift is thus the initial estimate of the velocity of the FG, i.e. the train. We denote this shift as v and this is our hypothesized train velocity in *pixel shifts/frame*. The correlation is computed by applying Normalized Cross Correlation (NCC) [15]. This technique is invariant to linear changes in illumination. A patch based correlation (and not a pixel) also ensures robustness to assumed Gaussian image noise.

Validation: Given an estimate v of the train velocity in terms of pixel shifts/frame, the next step is to validate other parts of the image and test if they conform to this motion. The regions which pass this validation test should correspond to FG, while the remaining regions will correspond to BG. But such a validation test will fail for texture-less (zero image gradient) BG regions, as two patches separated by some pixel-shifts/frame will match for any hypothesized velocity including v. To avoid this we incorporate the idea of validation to design a new cost function which can handle texture-less BG regions.

3.1. Generic Motion Estimation

In this section, we implement the validation step for each pixel and compute few quantitative values, which are later useful in designing our proposed cost function in Sec. 3.2. We

consider four image frames: I_{t-1}, I_t, I_{t+1} and I_{bg} (Fig. 2). Here, I_{bg} is the latest estimate of the BG image. The first instance of this image corresponds to the image frame captured just prior to the appearance of the train in the video. This is done by applying the Gaussian Mixture Model (GMM) based technique [4] to the image frames in the beginning of the video. As the BG is assumed to be visible at the beginning of the train video, this technique can model the BG quite efficiently and detect the train as a FG object as soon as it appears in the first image frame.

Now, lets consider a pixel p with coordinates (x, y) in I_t (Fig. 2). Its velocity is unknown to us. If p belonged to FG it should be observable at location $(x - v, y)$ in I_{t-1} and and at location $(x + v, y)$ in I_{t+1}. This is illustrated in Fig. 2 by the pixel surrounded by the dashed square window. We assume that a local patch around p also moves with velocity v and select square patches W_{t}, $W_{\text{t}-1}$, $W_{\text{t}+1}$ and W_{bg} centered at location (x, y), $(x - v, y)$, $(x + v, y)$ and (x, y) in image frames I_t, I_{t-1}, I_{t+1} and I_{bg} respectively.

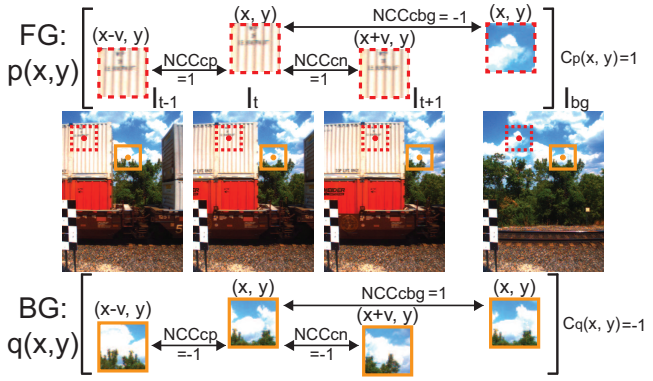

Fig. 2. Validating the hypothesized velocity v at two image patches: red (dashed boundary) belonging to FG and orange (solid boundary) belonging to BG.

Given $W_{\text{t}-1}, W_{\text{t}}, W_{\text{t}+1}$ and W_{bg}, we compute the NCC [15] values among these image patches as shown in Eq. 1,2,3. The NCC values lie between -1 and 1, where high value indicates matching candidate patches, while smaller values indicate unmatched patches. It can be observed that Eq. 1 and Eq. 2 encode the validation method (Sec. 3) as they check for the correctness of hypothesized velocity v using NCC_{cp} and NCC_{cn}. If they are close to 1, then W_{t} is a FG image patch. We denote these equations as *validation equations*. But, such a criteria alone will not detect texture-less BG correctly as shown in Sec. 3.2. The inclusion of NCC_{cbg} in the validation analysis will be critical to solve this problem. This forms the basis of our proposed cost function C_{p} in Eq. 4.

$$\text{NCC}_{\text{cp}}(x, y) = \text{NCC}[W_{\text{t}}(x, y), W_{\text{t}-1}(x - v, y)] \quad (1)$$

$$\text{NCC}_{\text{cn}}(x, y) = \text{NCC}[W_{\text{t}}(x, y), W_{\text{t}+1}(x + v, y)] \quad (2)$$

$$\text{NCC}_{\text{cbg}}(x, y) = \text{NCC}[W_{\text{t}}(x, y), W_{\text{bg}}(x, y)] \quad (3)$$

Before going further, we note that due to image distortion and perspective projection different parts of the train move with slightly perturbed values of v. Thus, we increase the set of hypothesized velocities to $v' = [v-\delta, v+\delta]$ and then compute Eq. 1-3 for each v'. We select the candidate with maximum value. We empirically set $\delta = 3$.

3.2. Proposed Cost Function

The problem with using simple validation based techniques (Sec. 3) and corresponding equations (Eq. 1,2) to classify $W_t(x, y)$ as FG/BG can be explained as:

Case 1. If $W_{\text{t}}(x, y) \in$ FG: $\text{NCC}_{\text{cp}} \approx 1$ and $\text{NCC}_{\text{cn}} \approx 1$ as we cross-correlate similar patches.

Case 2. If $W_{\text{t}}(x, y) \in$ BG: If BG is textured, then $\text{NCC}_{\text{cp}} \approx -1$ and $\text{NCC}_{\text{cn}} \approx -1$, but if BG is texture-less then $\text{NCC}_{\text{cp}} \approx 1$ and $\text{NCC}_{\text{cn}} \approx 1$ as the BG patches at (x, y),$(x-v, y)$ and $(x+v, y)$ are similar. This observation satisfies Case 1 above and classifies $W_t(x, y)$ as FG.

Case 3. If $W_{\text{t}}(x, y) \in$ FG+BG: If pixel p is located at the FG and BG boundary, then $W_{\text{t}}(x, y)$ will include both FG and BG regions. It is known that motion estimation in such regions is challenging [8]. In our case, we post-process these regions after our BG subtraction to get refined FG boundaries.

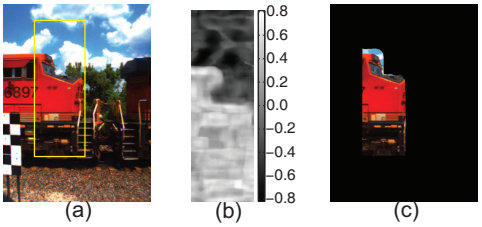

Fig. 3. (a) Input image. (b) FG cost C_{p}. (c) Extracted FG.

Based on these observations, we propose the following cost function using the NCC information available in Eq. (1,2,3):

$$\text{C}_{\text{p}} = [\text{NCC}_{\text{cp}} + \text{NCC}_{\text{cn}} - 2 * \text{NCC}_{\text{cbg}}]/4 \quad (4)$$

It can be observed that if $W_{\text{t}}(x, y) \in$ FG, we have $\text{C}_{\text{p}}(x, y) \approx 1$ and if $W_{\text{t}}(x, y) \in$ BG, then $\text{C}_{\text{p}}(x, y) \leq 0$ for textured as well as texture-less regions. The application of this cost function is demonstrated in Fig. 3 where BG needs to be subtracted in a block (yellow) at the center of the image. The BG consists of both textured (trees) as well as texture-less (blue sky, clouds) regions. The computed C_{p} at all pixel locations inside the block is shown in Fig. 3(b). As can be seen most of the BG have $\text{C}_{\text{p}} \leq 0$ (see color bar). To extract the FG inside the block, a threshold τ is set and each column of the cost C_{p} is compared to find the index where the threshold is reached. This gives the top edge of the container. From earlier analysis of C_{p}, ideally $\tau = 0$ should be able to differentiate BG and FG. In practice, we found that $\tau = 0.2$ gives better performance. This is because due to image noise there are no ideal texture-less regions and a slightly higher value of τ is preferable. Once the top boundary of the container is found, all the

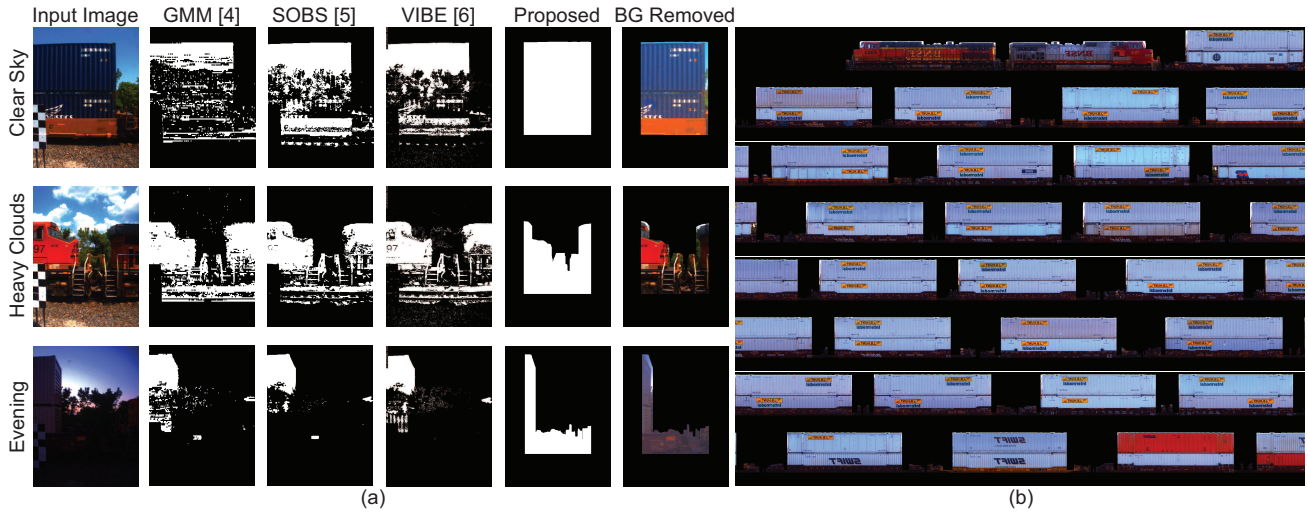

Fig. 4. (a) BG subtraction mask on different illumination conditions. The BG mask in proposed technique is missing at the boundaries as those regions were either not present or are lost in previous and next image frames respectively. (b) Orthographically projected panoramic mosaics. They can be concatenated from left to right to obtain the complete mosaic.

pixels below it are classified as FG (Fig. 3(c)). to construct panoramic mosaics as follows.

4. PANORAMIC MOSAIC GENERATION

Let a BG subtracted image corresponding to I_t be denoted as fg_t. Since we also know the pixel velocities for each pixel in the FG, they can be averaged to obtain a global velocity v_{fg} of the FG. This implies that a new image patch of width v_{fg} will be seen in frame I_{t+1}. Thus, one can select image patches of width v_{fg} from the center of each BG subtracted image (where image distortion is least) and concatenate such patches to create an orthographic mosaic of a BG subtracted IM train video. Such a technique guarantees that the shape of the train is not elongated or shortened due to overlap or underlap of image patches in the mosaic. A mosaic is useful for visualization of the complete train and can be used for classification of individual load types [2]. A sample mosaic for a train is shown in Fig. 4(b).

5. RESULTS

Data acquisition: A video acquisition system is installed at Sibley, MO, USA. A camera of focal length 8 mm is used to capture the video of the train at 30 fps. An auto-exposure routine is run before capturing a train and camera parameters are adjusted for current lighting conditions. These are then kept fixed for the entire video. The size of each image frame is 640×480. Each train consumes approximately $5 - 10$ GB of space depending on its length.

BG subtraction: The proposed technique has been tested on a wide variety of videos captured over a period of 12 months under varying illumination conditions. Our results of BG subtraction are shown in Fig. 4 on three different kinds of videos. The results are compared with three state of the art methods: GMM [4] (our implementation), SOBS [5] and VIBE [6] (author's implementation). We use all the implementation with default values and keep them same for all the experiments. Fig. 4(a) top row shows the results for a video with clear blue sky, where all the methods perform well. Although, the FG is classified as BG inside the containers as the texture of trees and container is similar. Fig. 4(a) middle row shows results for a cloudy sky. Here also the performance of our technique is at par with other methods. Although, the VIBE method is not able to remove all the BG. In Fig. 4(a) bottom row, we have a video captured during the evening when the illumination levels are really low. It can be seen that all the methods except for our technique fail to detect the FG near the bottom of the image (rail car and wheels).

Gap length accuracy: Fig. 4(b) shows the orthographic mosaic generated from BG subtracted images. The boundaries of the container have been post-processed to take care of error resulting from using patch-based technique for computing NCC values. This mosaic can be used to compute the length of all gaps in pixel lengths. We manually computed the accuracy of BG subtraction, by visually inspecting $22,000$ gaps in such mosaics and comparing the FG boundary from the video and the one computed in the mosaic. After defining an acceptable error threshold of ± 5 pixels, we computed the accuracy of BG subtraction to be 90.90%.

Computational speed: On a 2.67GHz, Intel Core i7 CPU with 64-bit windows, the BG subtraction and mosaic generation is done at the rate of 16.2 fps while the video acquisition rate is 30 fps.

6. CONCLUSION

This paper proposes a new cost function which can handle texture-less BG regions, while applying motion-based BG subtraction. It has been implemented as part of a machine vision system for analyzing gap lengths in IM freight trains. The system has been functional at an outdoor location.

7. REFERENCES

[1] Y-C Lai, C P L Barkan, J Drapa, N Ahuja, J M Hart, P J Narayanan, C V Jawahar, A Kumar, L R Milhon, and M Stehly, "Machine vision analysis of the energy efficiency of intermodal freight trains," *Proceedings of The Institution of Mechanical Engineers Part F-journal of Rail and Rapid Transit*, vol. 221, pp. 353–364, 2007.

[2] A. Kumar, N. Ahuja, J.M. Hart, U.K. Visesh, P.J. Narayanan, and C.V. Jawahar, "A vision system for monitoring intermodal freight trains," in *WACV*, Feb 2007, p. 24.

[3] Qing-Jie Kong, A. Kumar, N. Ahuja, and Yuncai Liu, "Robust segmentation of freight containers in train monitoring videos," in *WACV*, dec. 2009, pp. 1 –6.

[4] C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," in *CVPR*, 1999, vol. 2.

[5] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *Transactions on Imgage Processing*, vol. 17, no. 7, pp. 1168–1177, July 2008.

[6] O. Barnich and M. Van Droogenbroeck, "Vibe: A powerful random technique to estimate the background in video sequences," in *ICASSP*, april 2009, pp. 945 –948.

[7] A. Mittal and N. Paragios, "Motion-based background subtraction using adaptive kernel density estimation," in *CVPR*, 2004.

[8] S. Baker, D. Scharstein, J.P. Lewis, S. Roth, M.J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," in *ICCV*, oct. 2007, pp. 1 –8.

[9] M. Piccardi, "Background subtraction techniques: a review," in *International Conference on Systems, Man and Cybernetics*, Oct 2004, vol. 4, pp. 3099 – 3104.

[10] S. Brutzer, B. Hoferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *CVPR*, June 2011, pp. 1937 –1944.

[11] Christopher Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland, "Pfinder: Real-time tracking of the human body," *IEEE PAMI*, vol. 19, pp. 780–785, 1997.

[12] Zoran Zivkovic, "Improved adaptive gausian mixture model for background subtraction," in *ICPR*, 2004.

[13] Ahmed M. Elgammal, David Harwood, and Larry S. Davis, "Non-parametric model for background subtraction," in *ECCV*, 2000, pp. 751–767.

[14] Wei Li, Xiaojuan Wu, K. Matsumoto, and Hua-An Zhao, "Foreground detection based on optical flow and background subtract," in *Communications, Circuits and Systems (ICCCAS)*, july 2010, pp. 359 –362.

[15] JP Lewis, "Fast normalized cross-correlation," *Vision Interface*, vol. 10, pp. 120–123, 1995.