Disparity Map Refinement for Video Based Scene Change Detection Using a Mobile Stereo Camera Platform

Hakan Haberdar and Shishir K. Shah
Quantitative Imaging Lab
University of Houston
Houston, Texas, U.S.A.
hhaberdar@uh.edu, sshah@central.uh.edu

Abstract—This paper presents a novel disparity map refinement method and vision based surveillance framework for the task of detecting objects of interest in dynamic outdoor environments from two video stereo sequences taken at different times and from different viewing angles by a mobile camera platform. The proposed framework includes several steps, the first of which computes disparity maps of the same scene in two video sequences. Preliminary disparity images are refined based on estimated disparities in neighboring frames. Segmentation is performed to estimate ground planes, which in turn are used for establishing spatial registration between the two video sequences. Finally, the regions of change are detected using the combination of texture and intensity gradient features. We present experiments on detection of objects of different sizes and textures in real videos.

Keywords—change detection; texture; disparity map refinement; video surveillance; object detection; stereo vision

I. INTRODUCTION

The detection of relevant changes in video sequences of the same scene acquired at different times is crucial in vision tasks, especially in applications related to video surveillance, remote sensing, medical diagnosis, and driver assistance systems. The motivation of the work presented in this paper stems from video patrolling applications in outdoor environments, where the objective is to identify objects that may have been placed or removed from the scene to be monitored. Various vision based methods for detecting objects in a scene using stationary cameras have been reported in literature, while considerably fewer publications address the analysis of images acquired by mobile camera platforms [1]. Sand and Teller [2] describe an algorithm for bringing two videos into spatiotemporal alignment and compare them using feature point correspondences. The limitation of their approach is requiring the moving cameras to follow nearly identical trajectories. Primdahl et al. [3] present a method for automatic guidance of vehicles moving in well defined environment. This approach requires a priori understanding of the scene.

In this paper, we address the problem of detecting regions of change between two video sequences of the same scene taken by a mobile stereo camera platform at different times and different viewpoints. Using mobile platforms in outdoor environments has its own challenges such as complex image geometry, parallax, and variations in illumination conditions. The main contribution of this paper is twofold: the overall video surveillance framework and the corresponding algorithms for disparity map refinement and frame registration using ground planes in the presence of rigid and non-rigid motions, parallax problems, and changes in illumination conditions. In this work, we assume that temporal alignment between two video sequences is available. While the alignment does not provide for the same field-of-view, we assume that there is reasonable overlap to allow for the scenes to be compared. Our framework consists of five steps. It begins with a disparity map estimation module which is responsible for calculating depth information of the scene in two video sequences. These preliminary disparity maps are refined by using estimates from neighboring frames. To the best of our knowledge, there is no method utilizing sequential depth information for refining disparity maps in the manner proposed here. We then find the ground planes in each frame by comparing the texture of the disparity layers. Finally, we register the ground planes in two video sequences and compare them to detect regions of change using a combination of texture and intensity gradient features. We assume, in this work, that the objects or changes to be detected are on the ground.

The rest of the paper is organized as follows. Section 2 gives an overview of the framework. Section 3 discusses experimental results. Finally, section 4 concludes the paper.

II. VIDEO SURVEILLANCE FRAMEWORK

The proposed video surveillance framework (shown in Fig. 1) assumes the existence of temporal alignment between video sequences acquired before and after the introduction of changes in the scene. We accomplish spatial registration and change detection in five steps: (1) disparity estimation in the before and after stereo images; (2) refinement of the individual disparity estimates using sequential stereo images; (3) ground plane segmentation using texture and disparity estimates; (4) spatial transform estimation using ground plane; and (5) detection of textural anomalies to delineate areas of change. Within this framework, no a priori
information about the changes in the scene and trajectory of the platform is assumed, except that the changes to be detected are at the ground level. In the following, we present details of each of the steps in the proposed framework.

A. Estimation of Disparity Map

In Fig. 1, V1 and V2 denote 2 stereo video sequences, V1 : (Fr1R, Fr1L), ..., (FrN_R, FrN_L) and V2 : (Fr1R, Fr1L), ..., (FrN_R, FrN_L), ..., (FrM_R, FrM_L) where Fr1R and Fr1L are right and left images of the stereo frame i, and the image pairs Fr1R & Fr1L and FrN_R & FrN_L show similar scene content. The goal of our stereo matching module is to estimate disparity maps D1 and D2 using image pairs Fr1R & Fr1L and Fr2R & Fr2L, respectively. Each disparity map will provide a representation of the depth of the scene. Existing stereo matching algorithms can be categorized into two classes. The first class, window-based algorithms, use intensity values within a finite neighboring window to determine the disparity at a given pixel. They produce noisy disparities in textureless regions and discontinuous disparity boundaries [4]. The second class is global algorithms where smoothness assumptions of the disparity map are made, and various minimization techniques are employed. Graph cuts, one of the global methods, recently has attracted much attention due to its optimality properties and the success of reported results. Accordingly, we use the graph cut based energy minimization technique [5] to compute disparity maps D1 and D2.

B. Refinement of Disparity Map

Due to the ill-posed nature of the stereo matching problem, estimated disparity maps D1 and D2 sometimes exhibit very high level of noise. We propose alleviating noisy estimates based on a refinement process. Based on the assumption that consecutive disparity maps in a video sequence cannot change dramatically, we can examine whether estimated disparity values in the sequential frames are reasonable and consistent. In our refinement strategy, we perform the following two controls: point based tracking and layer based consistency. First of all, our algorithm selects ordinary control points in the horizontal and vertical direction separated by 3 pixels forming a fine grid in frame Fr1L, and then it tracks those points across four consecutive frames Fr1_L, Fr1_L, Fr1_L, and Fr1_L using the Lucas-Kanade tracker [6]. Let dP denote the disparity value of a control point p in the disparity map D1, and let level_threshold denote threshold that defines the acceptable level of change in disparity between two consecutive frames. If |dP−dP−1| < level_threshold, |dP−dP−2| < level_threshold, and |dP−dP−2| < level_threshold, and |dP−dP−2| < level_threshold, we do not change the disparity value dP. If the disparity value of point p in the disparity maps D1, D1−1, D1+1, and D1+2 does not satisfy the above conditions, we use a disparity level mean function to estimate the new disparity value of point p. Let dP_new = meanlevel(dP−2,dP−1,dP+1,dP+2). Following this, the 8 neighboring pixels of the point p are set to the same disparity value dP_new. Finally, we use a modified median filter to overcome problems due to noisy points and guarantee that adjacent control points have similar disparity values while preserving edges. The limitation of our refinement strategy is that it requires at least two disparity values out of five to be calculated accurately. Hence, before we start the refinement process, we search the entire disparity map sequence until we find reliable disparity values. In this work, we have set the value of level_threshold to be 30 where normalized disparity values range from 0 to 255 (e.g., 0 refers to the furthest point).

The layer based consistency relies on the idea that size of a disparity layer in consecutive frames cannot change dramatically. We expect smooth disparity value changes of a layer since a layer may come closer to (or move away from) the camera in a consecutive frame. Let sLi denote size of a disparity layer L in the disparity map D1, and let size_threshold denote threshold specifying the degree of reasonable change in size. If ∑ j=−2,−1,1,2 (sL+1)2 + (sL)2 > size threshold, we restore the layer L in the disparity map D1 using four disparity maps in the sequence. The size_threshold is set to allows for a 20% change in the size of the layer being considered.

We iterate point and layer based controls until there are no more changes in the refined disparity map D1. We perform all the same steps for the disparity map D2 to obtain enhanced disparity map D2. Fig. 2 shows the effect of disparity refinement on ground plane estimation. Building the framework on accurately refined disparity maps allows robust ground plane estimation and avoids of many of the problems related to the spatial registration.
C. Ground Plane Estimation and Region Merging

Disparity map refinement is followed by a layer-based segmentation step allowing for the estimation of the position of the ground plane within each image. We represent the scene structure as a collection of disparity layers. Many stereo algorithms impose the assumption that all 3D points in each layer lie on the same plane in the 3D world and the disparities in each layer obey the same plane equation, whereas resulting disparity maps sometimes include over-segmented layers. We relax the constraint and assume that the extracted disparity layers do not adhere to a single scene layer. In contrast, we allow for a region merging step that can analyze the estimated disparity maps to generate a refined layer representation (Fig. 3).

We start by constructing a graph on the over-segmented disparity map, as shown in Fig. 3, whose nodes represent the candidate disparity layers and an edge refers to the first order neighborhood relationship of a candidate layer. The first disparity layer (i.e., the node 0) is assumed to be ground plane, knowing that other candidate layers may also be part of the ground plane. In many cases, there may also be more than one root node. We employ a log-Gabor filter bank to extract texture features of each candidate disparity layer [7]. Next, we compute the median absolute deviation (MAD) [8] of each layer using the filter bank output. We define the metric $m_{MAD} = \frac{(MAD_1)^2 + (MAD_2)^2}{MAD_1 * MAD_2}$, where MAD_1 and MAD_2 are MADs of the layers to be compared. We perform a pair-wise comparison of MAD values of neighboring layers. The layers L_i and L_j are merged if the metric $m_{MAD}$ is less than $MAD_{threshold}$. The value of $MAD_{threshold}$ is fixed to be 2.25, where a value of 2.0 indicates that two layer have exactly the same texture.

D. Registration of Ground Planes

Many change detection methods utilize an entire image region for spatially registering two images, imposing the same transformation function for a whole scene. This is normally not true since outdoor scenes often include different planes and exhibit parallax. Therefore, we use only the ground planes in the registration step, especially knowing that the changes to be detected are also on the ground. Specifically, to register the two ground planes, we estimate the parameters of the homography transform. Ground plane

E. Change Detection

In the change detection step, the goal is to distinguish changed pixels from the ground plane pixels. Illumination distributions of the scene in the images $G^1_{639}$ and $G^2_{1057}$ (Fig. 4) are different because of temporal separation, and
there are also large shadow regions within the images. Therefore, instead of utilizing intensity values and performing basic frame differencing, we divide $G_{1057}^{2}$ and $G_{639}^{2}$ into a number of square regions (Fig. 4) and compute texture features of each sub-region. Texture representations are more accurate than local statistics, but when regions of change and the scene are homogeneous, the texture difference measure will fail. Therefore, we evaluate the integrated gradient values for change detection and calculate image intensity gradients of corresponding sub-regions in two images. If the absolute difference of MAD values for combined texture and gradient features of the corresponding squares in $G_{639}^{2}$ and $G_{1057}^{2}$ is greater than $\text{MAD}_{\text{Tex-Gr}}^{\text{threshold}}$, we set the square region as a region of change. Fig. 4 shows an example of change detection where the box is detected successfully.

III. EXPERIMENTAL RESULTS

We have evaluated our framework using 4 video sequence pairs that were acquired by a mobile stereo camera platform in real outdoor environments under differing illumination conditions. For each environment, we have 2 videos called before and after, where after was taken after objects of different sizes and textures were placed in the scene, and before was taken without the objects. Each video is approximately one and half minutes (i.e. $\sim 3000$ frames) in length. Table I shows the number of frames in which an object can be seen and the number of frames where we detected the object successfully. Results of detecting the change with and without disparity map refinement are presented and clearly show the advantage of refinement.

### Table I

<table>
<thead>
<tr>
<th>Object ID</th>
<th>Viewable</th>
<th>Detected</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>original</td>
<td>refined</td>
<td>original</td>
</tr>
<tr>
<td>1</td>
<td>43</td>
<td>26</td>
<td>34</td>
</tr>
<tr>
<td>2</td>
<td>61</td>
<td>23</td>
<td>51</td>
</tr>
<tr>
<td>3</td>
<td>57</td>
<td>45</td>
<td>53</td>
</tr>
<tr>
<td>4</td>
<td>68</td>
<td>53</td>
<td>59</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>21</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>285</td>
<td>179</td>
<td>242</td>
</tr>
</tbody>
</table>

For objects 1, 5, and 7, our change detection module failed for some of the frames due to problems related to spatial registration. For objects 2 and 4, most of the failure were attributed to inaccurate estimation of disparity maps. For objects 3 and 6, our ground plane estimation sometimes produced incorrect results such as labeling closer objects as part of the ground plane. Nevertheless, since this was the same case for both frames before and after, the texture comparison module was successfully able to detect the objects.

We performed a second experiment to observe the accuracy of the proposed framework when we do not place any objects in the scene. This was based on analyzing randomly picked 60 consecutive frames where there were no objects introduced in the scene. Our framework could recognize that there was no change in 53 of the frames (i.e., 88.33% accuracy). Overall, the framework could detect changes with reasonable accuracy.

IV. CONCLUSION

In this paper, we have presented a novel disparity map refinement method and video surveillance framework for detecting objects in outdoor environments using disparity maps from a mobile stereo camera platform. The framework is evaluated on 4 real video sequences and the experimental results show that the proposed framework is able to detect changes/objects of different sizes and textures, and is able to ascertain if there are no changes in the scene, thereby minimizing false positives. Our future work will focus on relaxing our assumption on known temporal alignment and improving the spatial registration module.

ACKNOWLEDGMENT

We would like to thank Mr. Jayan Eledath of Sarnoff Corp. for his invaluable support and guidance through this work. The research presented in this paper is based upon work supported by, or in part by, the U. S. Army Research Office under contract/grant number W911NF-08-0326.

REFERENCES


