Fast Illumination-invariant Background Subtraction using Two Views: Error Analysis, Sensor Placement and Applications

Ser-Nam Lim[‡] Anurag Mittal[§] Larry S. Davis[‡] Nikos Paragios[£*]

CS Dept., University of Maryland, College Park[‡] Siemens Corporate Research, Princeton, NJ[§] Ecole Nationale des Ponts et Chaussees[£]

Abstract

Background modeling and subtraction to detect new or moving objects in a scene is an important component of many intelligent video applications. Compared to a single camera, the use of multiple cameras leads to better handling of shadows, specularities and illumination changes due to the utilization of geometric information. Although the result of stereo matching can be used as the feature for detection, it has been shown that the detection process can be made much faster by a simple subtraction of the intensities observed at stereo-generated conjugate pairs in the two views. The method, however, suffers from false and missed detections due to some geometric considerations. In this paper, we perform a detailed analysis of such errors. Then, we propose a sensor configuration that eliminates false detections. Algorithms are also proposed that effectively eliminate most detection errors due to missed detections, specular reflections and objects being geometrically close to the background. Experiments on several scenes illustrate the utility and enhanced performance of the proposed approach compared to existing techniques.

1 Introduction

Foreground object detection using background subtraction (e.g. [1, 2, 3]) has been used extensively in video surveillance applications due to its underlying ease of implementation and effectiveness. Most previous work has focused on using a single camera for background modeling, which is highly effective for many common surveillance scenarios. However, it is difficult to deal with sudden illumination changes and shadows when only a single camera is used.

The use of two cameras for background modeling serves to overcome these problems. In particular, dense stereo correspondence between two views can be used to create a disparity map, which is invariant to shadows and illumination changes. Such a disparity map can be used as an input to a disparity-based background model, in principle achieving robustness against illumination changes.

Since it is necessary that accurate stereo correspondences be used for the background model (e.g. [4]), sophisticated stereo algorithms such as those described in [5, 6] can be used. However, without the aid of specialized hardware, most of these algorithms perform too slowly for real time background subtraction. Consequently, in many systems, the online stereo algorithm is implemented on hardware and is based on simpler and less accurate stereo. Examples include the SVM system described in [7, 8] and the video-rate stereo machine described in [9]. In [10]; disparity-based background modeling was similarly implemented on a single PCI card. This is then combined with color background subtraction so that foreground objects close to the background can be reliably detected.

1.1 Fast Illumination-Invariant Background Modeling using Multiple Views

Ivanov et. al. [11] described a clever method that does not require any specialized hardware but yet performs at video-rate. It employs accurate stereo to construct the background model, but rather than performing online stereo and disparity differencing for detection, the color difference between conjugate pixels is used to distinguish between background and foreground. Assuming that the scene is Lambertian and that the images have been color calibrated, the intensities for both pixels of a conjugate pair will change in the same way if they both view the background (which may become shadowed or illuminated differently), but differently if only one of them is the image of a foreground object. By utilizing disparity information implicitly, this method retains the advantages of multiple-view detection (invariance to illumination changes and shadows) while being very fast (≈ 25 fps). Since stereo is performed offline for background modeling, accurate stereo algorithms can be employed.

The algorithm inherently suffers from both missed and false detections (occlusion shadows) generated by homogeneous foreground objects. [11] suggested using additional cameras to mitigate the false alarms caused by occlusion shadows, but did not discuss how to reduce missed detections. Sensor placement, which affects these online error rates, was also not addressed.

In this paper, we analyze the problems of missed and false detections in Sec. 2. We describe an approach to the false detection problem from a sensor planning perspective in Sec. 3. In particular, we apply the algorithm from [11] using a two-camera configuration, in which the cameras are

*Author was involved with this work while he was at Siemens Corporate Research, Princeton, NJ.

vertically aligned w.r.t. a dominant ground plane i.e. the baseline is orthogonal to the plane on which foreground objects will appear. This configuration provides an initial foreground detection free of false detections. By sampling a small number of pixels from this initial foreground detection and generating stereo matches for them, we show that the missed detections can then be reduced in Sec. 4. Since only a small number of online stereo matches is required, system performance is not compromised. Sec. 5 concludes the paper with experimental results.

2 A Geometric Analysis of Missed and False Detections

2.1 False Detections (Occlusion Shadows)

Given a conjugate pair (p, p'), false detection of p occurs when p' is occluded by a foreground object but p is not. Ivanov et. al. [11] suggest the use of multiple cameras for this problem, detecting a change only when the difference from all of the other cameras is above a threshold. This idea, however, should be combined with proper sensor planning so that neighboring occlusion shadows as well as neighboring correct and missed regions do not overlap.

Consider a foreground object. We define three tangent points on the object as shown in Fig. 1(a): t_{ref} corresponds to the leftmost tangent line from the reference view[1], t_{sec_1} and t_{sec_2} correspond to both tangent lines from the second view respectively. Also, let the background pixels corresponding to them be b_{ref}, b_{sec_1} and b_{sec_2} respectively. Clearly, these points depend on the baseline, object size and object position. The extent E_p of the region of false detection is:

$$E_p = \min(||Pb_{sec_1} - Pb_{sec_2}||, ||Pb_{ref} - Pb_{sec_2}||), \quad (1)$$

where P is the projection matrix of the reference camera.

2.2 Missed Detections

Missed detections occur when a homogenous foreground object occludes both pixels of a conjugate pair, since the two pixels will then be very similar in intensity.

A simple geometrical analysis reveals that the extent E_n of the region of missed detection is dependent on the baseline, object size and object position. Referring to Fig. 1(a), E_n can be expressed as:

$$E_n = \max(||Pb_{sec_1} - Pb_{sec_2}|| - ||Pb_{ref} - Pb_{sec_2}||, 0). \quad (2)$$

[1]The reference view is the image in which we identify foreground pixels;clearly either of the two images can serve as reference.

As the distance between a foreground object and the background decreases, E_n approaches the extent of the image of the object. Thus, when the foreground object is sufficiently close to the background, it is entirely missed. This is a common problem associated with disparity-based methods, as mentioned earlier.

Eqns. 1 and 2 suggest that there is a tradeoff between the extent of false and missed detections that depends on the placement of the sensors. Thus, one can select the sensor placement that yields the desired trade-off. This is considered in the next section. The algorithm from [11] was tested on a real scene in Fig. 2. One can clearly see both missed and false detections.

3 Sensor Placement to Eliminate False Detections

In most surveillance applications, the objects (e.g. people and cars) to be detected are standing and moving on a dominant principle plane, which we refer to as the ground plane. For such applications, we consider a two-camera configuration that is well suited for dealing with false detections. The two cameras are placed such that their baseline is orthogonal to the ground plane and the lower camera is used as the reference for detection [Fig. 1(c)]. In this camera configuration, the epipolar planes are orthogonal to the ground plane.

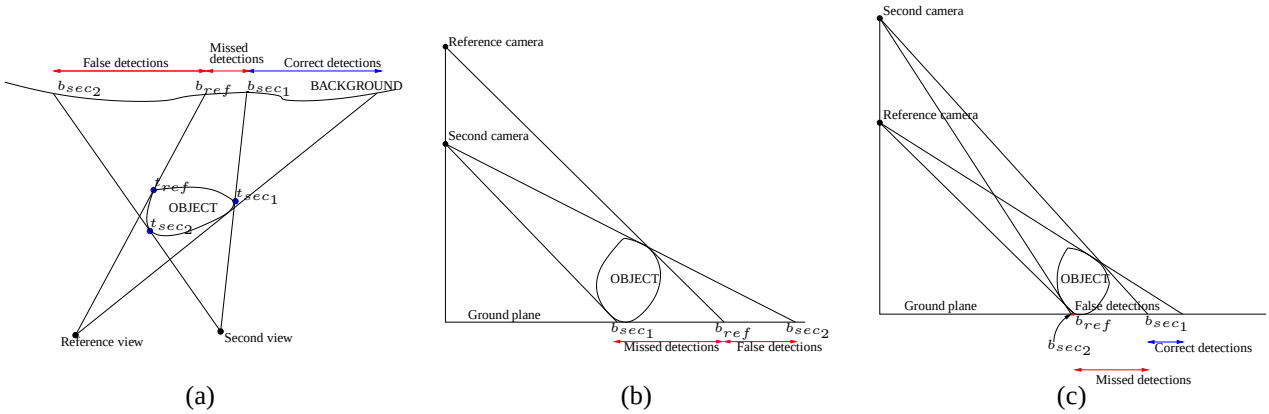
From Fig. 1(c), one can observe that if the lower camera is used as the reference, false detections can only be generated at the lower edge (edge closest to the ground plane) of the object. This is as opposed to using the higher camera as reference, shown in Fig. 1(b). Since objects are on the ground plane, E_p in Eqn. 1 is close to zero, in effect eliminating any false detection. Additionally, false detection does not occur at the left or right edge since the epipolar planes are orthogonal to the ground plane. Such a sensor configuration will be used throughout the rest of the paper.

On the other hand, missed detections remain at the lower portion of the object [Eqn. 2]. However, for an upright object that has negligible front-to-back depth, it may be shown (see Proposition 1 below) that the proportion of an object that is missed is invariant to its position. This result will play an important role in eliminating missed detections.

We assume that foreground objects are homogeneous, that the background pixels arise from the ground plane, and that objects are upright w.r.t. the ground plane. Then it is easy to show that:

Proposition 1 *In 3D space, the missed proportion of an homogeneous object with negligible front-to-back depth is independent of object position. Equivalently, the proportion that is correctly detected remains constant.*

Proof Referring to Fig. 3(a), the height of the object is h and that of the second camera is H. Let the length of the baseline be ℓ_b. The extent of the region of missed detection



(a) (b) (c)

Figure 1: Problem with the algorithm from [11]. (a) Missed and false detections shown from the top view. (b) Analysis for the special case of cameras vertically aligned w.r.t. the ground plane (side view). Here the top camera is taken as reference, which causes missed detection of the whole object and false detections as shown. (c) Switching the reference camera to the lower one eliminates most of the false detections, but missed detections remain according to Eqn. 2.

is $h - \frac{z_2 - z_1}{z_2}\ell_b$, thus giving the proportion ρ of the object that is missed as:

$$\begin{aligned} \rho &= \frac{h - \frac{h}{H} * \ell_b}{h} \\ &= 1 - \frac{\ell_b}{H}. \end{aligned} \tag{3}$$

Consequently, ρ is a constant, independent of the location of the object on the ground plane. □

Ideally, one would like to place the reference camera as close to the ground plane as possible so that ρ becomes zero. This is clear from Eqn. 3, where a baseline of length H eliminates any missed detection. However, mounting limitations, occlusion considerations and the imaging resolution of the ground plane typically limit the maximum possible length of the baseline, leaving some missed detections at the bottom of the object. Moreover, for outdoor scenes, it is clearly necessary that the reference camera be above the object so that the corresponding background is well-defined.

The usefulness of such a sensor configuration is illustrated in Fig. 4. Sudden illumination changes caused by a vehicle's headlight are detected when single camera background subtraction is used. On the other hand, by simply using the algorithm from [11] with the proposed sensor configuration, the detection results are invariant to the illumination changes while false detections are effectively prevented.

4 Reducing Missed Detections

Using the proposed sensor configuration, an initial detection generally free of false detections can be obtained; missed detections however remain at the lower portion of each object. In this section, we describe an approach to reduce these missed detections.

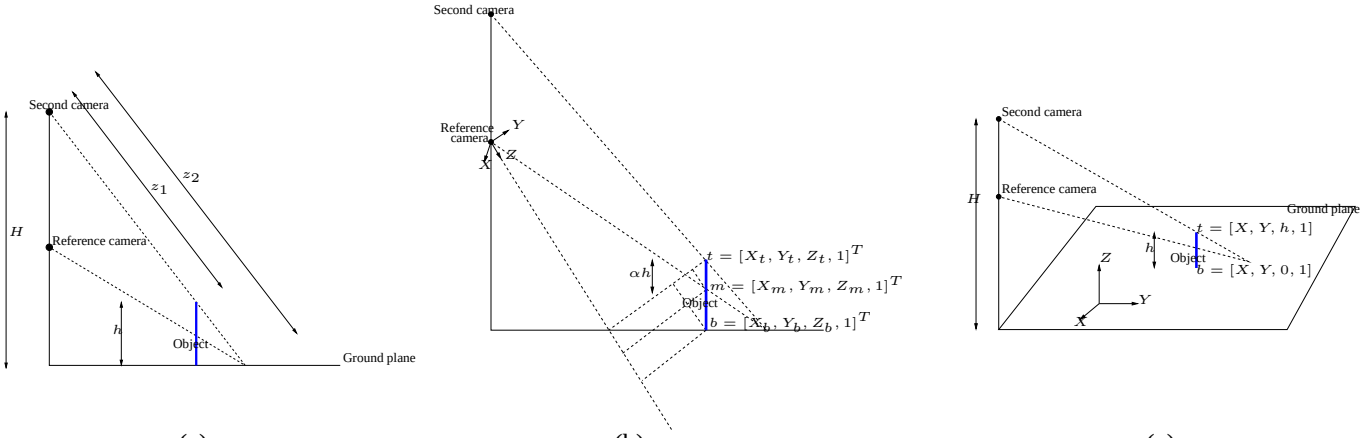
Let ω be a foreground blob from the initial detection, and let I_t be a foreground pixel in ω with its corresponding 3D point being t. Define the base point, b, of t as the point on the ground plane below t. The image of b is denoted as I_b.

A stereo search, constrained to only foreground pixels in the second view lying along the associated epipolar line, is first used to find the conjugate pixel $I_{t'}$ of I_t. The location of I_t and $I_{t'}$, together with calibration information allows us to determine I_b, as described in Sec. 4.1.

If $||I_t - I_b||$ is sufficiently large, then I_t is an off-ground-plane pixel and we begin a search along the epipolar line through I_t to find the location where the ground plane is first visible. We employ an iterative approach that works as follows: we first increment I_t by ΔI_t along the associated epipolar line, and then the base point, I_b, for the new I_t is determined in the same fashion. When $||I_t - I_b||$ is less than some critical value, then we have found the lower boundary of the foreground blob along the associated epipolar line.

ΔI_t must lie in the interval $[1, ||I_t - I_b||]$ pixels. Using the lower bound for ΔI_t generally gives a well-defined foreground blob, while using the upper bound generates a foreground bounding box. The trade-off is the number of stereo searches, decreasing as ΔI_t increases. The algorithm can also be easily extended to handle objects not moving on the ground plane surface. In this case, the iteration is terminated when the base points of the sampled pixel and the corresponding background are sufficiently close.

Figure 2: Detection results using the algorithm from [11]. Left to right: reference view, second view and foreground detections. Both missed and false detections are clearly evident.



(a) (b) (c)

Figure 3: (a) The proportion of missed detections for a homogeneous object with negligible front-to-back depth is independent of object position. (b) Image projection in camera-centered 3D coordinate system. (c) Image projection with 3D coordinate system on the ground plane.

Figure 4: Left to right: reference view, second view, single camera detection and two-camera detection. The usefulness of the proposed sensor configuration is illustrated here. The vehicle's headlight is cast on the wall of the building.

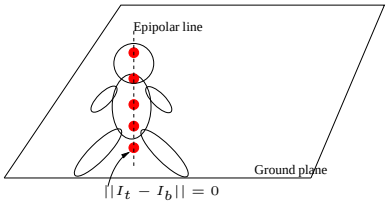


Figure 5: Along the epipolar line, the red dots are the sampled pixels. The lowermost sampled pixel has $||I_t - I_b|| = 0$ since it lies on the ground plane and is consequently used as the lower boundary of the foreground blob along the epipolar line.

4.1 Determining the Base Point

The algorithm requires that the base point of a pixel be determined. This can be achieved with two different approaches. The first approach assumes weak perspective projection i.e. all points on an object have the same depth. This is often a good approximation for outdoor scenes where objects are relatively far from the cameras. When this assumption is not valid, a second approach can be considered that utilizes the vertical vanishing point and the vanishing line of the ground plane. The details of both approaches are discussed in Sec. 4.1.1 and Sec. 4.1.2 respectively.

4.1.1 Weak Perspective Model

We use a camera-centered 3D coordinate system as shown in Fig. 3(b). t is the corresponding 3D point of I_t. Let its 3D coordinate be $[X_t, Y_t, Z_t, 1]$. The point m with 3D coordinate $[X_m, Y_m, Z_m, 1]$ is defined as the point such that its image I_m has coordinate $\Pi^{-1} * I_{t'}$, where Π is the ground plane homography from the reference to second view and $I_{t'}$ is the conjugate pixel of I_t in the second view. b is the base point of t with 3D coordinate $[X_b, Y_b, Z_b, 1]$. Let its image be I_b. We will first consider image projection in the $y-$direction. From the property of similar triangles, it can easily be verified from Fig. 3(b) that:

$$\frac{Y_t - Y_m}{Y_t - Y_b} = \alpha. \tag{4}$$

Here, $\alpha = 1 - \rho$ [Eqn. 3]. Consequently, Y_m and Y_b can be expressed as:

$$Y_m = Y_t - \alpha Y_t + \alpha Y_b, \tag{5}$$

$$Y_b = Y_t - \frac{Y_t}{\alpha} + \frac{Y_m}{\alpha}. \tag{6}$$

The image positions y_t, y_m and y_b of Y_t, Y_m and Y_b respectively can be expressed as:

$$y_t = Y_t \frac{f}{Z_t}, \tag{7}$$

$$y_m = Y_t \frac{f}{Z_m} - \alpha Y_t \frac{f}{Z_m} + \alpha Y_b \frac{f}{Z_m}, \tag{8}$$

$$y_b = Y_t \frac{f}{Z_b} - Y_t \frac{f}{\alpha Z_b} + Y_m \frac{f}{\alpha Z_b}, \tag{9}$$

f being the focal length. We are interested in the image ratio $\frac{||y_t - y_m||}{||y_t - y_b||}$. In the weak perspective case, the depths of points on the object are assumed to be a constant Z_{ave}. This gives:

$$\begin{aligned} \frac{||y_t - y_m||}{||y_t - y_b||} &= \frac{\alpha \frac{f}{Z_{ave}}(Y_t - Y_b)}{\frac{f}{\alpha Z_{ave}}(Y_t - Y_m)}, \\ &= \alpha. \end{aligned} \tag{10}$$

This shows that the detection ratio is an invariant under weak perspective assumption. The same principle applies to the image projection in the $x-$direction. Thus, using L_2 norm, $||I_t - I_m|| = \sqrt{\alpha^2(||y_t - y_b||^2 + ||x_t - x_b||^2)}$ and $||I_t - I_b|| = \sqrt{(||y_t - y_b||^2 + ||x_t - x_b||^2)}$, giving the detection ratio $\frac{||I_t - I_m||}{||I_t - I_b||} = \alpha$. Consequently, I_b is given as:

$$I_b = I_t + \frac{||I_t - I_m||}{\alpha}. \tag{11}$$

Notice that I_m can be determined independently using Π and $I_{t'}$. As a result, previous assumptions made in Eqn. 3 that the object is homogeneous and the background pixels are lying on the ground plane are unnecessary.

4.1.2 Perspective Model

When the weak perspective assumption is violated, the base point can be better estimated by using additional image-based calibration information in the form of the vertical vanishing point and the vanishing line of the ground plane. This method is based on the work of Criminisi et. al. [12], who described a method for computing distances between parallel planes in a single view. In particular, we extend their method to determine the image of the base point for a conjugate pair. It may be noted that the calibration required for this approach is simpler than full camera calibration required for Euclidean reconstruction.

Consider the projection matrix P of the reference camera. Let $[P_1 \;\; P_2 \;\; P_3 \;\; P_4]$ represents its matrix columns. The 3D coordinate system is as shown in Fig.3(c). Consequently, let the 3D coordinates of t and b be $[X, Y, h, 1]^T$ and $[X, Y, 0, 1]^T$ respectively, where h is the height of the object above the ground plane. The images of t and b can thus be expressed as:

$$\begin{aligned} I_b &= \beta_b(XP_1 + YP_2 + P_4), \\ I_t &= \beta_t(XP_1 + YP_2 + hP_3 + P_4). \end{aligned} \tag{12}$$

β_b and β_t are unknown scale factors. Let the normalized vanishing line and vertical vanishing point of the ground plane be $\hat{\ell}_{ref}$ and v_{ref} respectively. Since P_3 is actually the vertical vanishing point scaled by an unknown factor β_{ref}, the following is true:

$$I_t = \beta_t(\frac{I_b}{\beta_b} + h\beta_{ref}v_{ref}). \tag{13}$$

If we take the vector product of both terms of Eqn. 13 with I_b, followed by taking the norm on both sides, the following expression results:

$$h\beta_{ref} = -\frac{||I_b \times I_t||}{(\hat{\ell}_{ref}.I_b)||v_{ref} \times I_t||}. \tag{14}$$

The above derivation was first presented in [12]. β_{ref} can be computed if we know the height of a reference object in the scene. Due to errors present in the computation of $\hat{\ell}_{ref}$ and v_{ref}, it is often required that more robust methods be used for computing them. In [12], a Monte-Carlo approach was used.

The same principle can also be applied to the second camera. Let the parameters for the second camera be β_{sec}, $\hat{\ell}_{sec}$ and v_{sec}. Consequently, we can equate the height in Eqn. 14 for both cameras to obtain the following equation:

$$\frac{||I_b \times I_t||}{\beta_{ref}(\hat{\ell}_{ref}.I_b)||v_{ref} \times I_t||} = \frac{||(\Pi * I_b) \times I_{t'}||}{\beta_{sec}(\hat{\ell}_{sec}.(\Pi * I_b))||v_{sec} \times I_{t'}||}. \tag{15}$$

The image of the base point in the second view is clearly $\Pi * I_b$, where Π is the ground plane homography. $I_{t'}$ is again the conjugate pixel of I_t. In addition, I_b is constrained to lie on the line through I_t and the vertical vanishing point. I_b can thus be computed using these two constraints.

5 Implementation and Results

Experiments were performed on a dual Pentium Xeon, 2GHz machine. We utilized the extra processor to perform in parallel single camera background subtraction in the second camera. The resulting performance of the system was very fast, with frame rate in the range of ≈ 25 fps.

Correspondences of background pixels for the background model were mainly determined using homographies of the principle planes present in the scene, computed on the basis of a small set of manually selected matches per plane. This typically leaves only a small set of background pixels for general stereo matching. Background subtraction is performed by computing the normalized color difference for a background conjugate pair and averaging the component differences over a $n \times n$ neighborhood (typically 3×3). To deal with different levels of variability, each background conjugate pair is modeled with a mixture of Gaussians ([2]) that are updated over time (typically two Gaussians are sufficient). Foreground pixels are then detected if the associated normalized color differences fall outside a decision surface defined by a global false alarm rate.

While the two-camera algorithm will not detect shadows as foreground, it will generally detect reflections of the foreground objects from specular surfaces, such as wet pavement, as foreground. We describe below a simple method that removes most of these specular reflections.

First, after applying the basic two-camera algorithm we employ simple morphology and connected component analysis to find foreground objects. This is illustrated in Fig. 6(a), 7(a), 8(a) and 9(a), where we show the bounding boxes that surround these foreground pixel clusters detected by this spatial clustering step.

Employing our base-finding algorithm, we first find the intersection of the foreground object with the ground plane as follows. The "topmost" pixels of the foreground region along each epipolar line passing through the bounding box are identified, and for each of these topmost pixels we evaluate the image gradient to determine whether they are good candidates for stereo matching. This will typically choose those pixels on the boundary of the object detected. For each of those pixels, we find the base using the algorithm from Sec. 4.1.1. The line passing through the bases can then be constructed using a robust line fitting algorithm. The object is detected by "filling in" the foreground region above the base line along the epipolar lines. This is illustrated in Fig. 6(c), 7(c), 8(c) and 9(c). The first step in finding the base is identifying the conjugates for these topmost points; to make this efficient, we constrain the matches to only those pixels in the second view along the epipolar line that are additionally foreground pixels detected by a single camera background subtraction algorithm (which will detect a superset of the pixels detected by the two-camera algorithm). The results of the single camera background subtraction applied to the second view are shown in Fig. 6(b), 7(b), 8(b) and 9(b).

As a side effect, we can eliminate from the initial detection any pixel detected as a foreground pixel but lying below the base of the object. This tends to eliminate specular reflections "connected" to the foreground region by the spatial clustering step. The reason is that the virtual image of an object reflected from the ground plane lies below the plane. However, it is possible that a component of reflected pixels in the reference image is not connected by the spatial clustering algorithm to the object that cast the reflection. In this case, we find that, typically, the stereo reconstruction algorithm fails to find good matches along the epipolar line in the second view. This is not surprising since the observed input results from a combination of Lambertian and specu-

lar components at the point of reflection. The likelihood of getting a match is low because a difference in either the Lambertian components or the reflection properties would cause the reflected points to appear differently. Even if they are matched, the base point would lie above the reflected point. Thus, we typically eliminate these specular components also. This can be seen in Fig. 6(c) and 7(c) - notice the bounding box below the vehicle in Fig. 7(b). It is a specular reflection from the vehicle, and is eliminated due to failure to find conjugates in the second view. Many more examples of the elimination of specular reflections can be seen in the accompanying videos.

A common problem associated with disparity-based background subtraction occurs when the foreground object is physically close to a surface such as a wall of a building. Gordon et. al. [10] proposed a solution to this problem that combines disparity and color information. However, since disparity information for the whole image is required, performance can become a concern. Furthermore, although the method utilizes adaptive thresholding, it is not fully invariant to shadows and illumination changes. On the other hand, because our algorithm requires only initial partial detection, its performance in detecting near-background objects compares favorably. In particular, when a foreground object comes close to a background surface such as a wall, the algorithm can typically still detect the top portion of the object. This initial detection can subsequently be used to initialize our base-finding algorithm. We demonstrate this in Fig. 8. Besides some specularities (reflection in the long glass windows) and shadows (on the wall), the person was also walking near the background wall. In spite of that, the person was fully detected without any false alarms.

The perspective model is important for indoor scenes, where objects are closer to the camera. An example is shown in Fig. 9, where in (e), the bases of three chosen pixels are used to form the lower boundary of the object. Comparison with the weak perspective model is also shown in (d). With accurate calibration, the perspective model also performs as well as the weak perspective model for outdoor scenes. For example, the perspective model was used to compute the base point in Fig. 8.

6 Concluding Remarks

This paper considers a fast background subtraction algorithm using two cameras that has been previously considered in the literature [11]. This algorithm has the advantage of being extremely fast and simple while being invariant to shadows and illumination changes. However, the application of the method results in both false and missed detections due to certain geometric considerations. In this paper, we have analyzed these errors in terms of the camera geometry. From the analysis, a sensor configuration was proposed that effectively eliminates most false detections. Additionally, algorithms were considered that fill-in missed detections and eliminate false detections occurring as a result of specularities. The result is a surveillance system that gives very accurate detection in an extremely efficient manner without significant errors due to shadows, sudden illumination changes and specularities. Due to these characteristics, the system can be very useful in surveillance applications where high performance is critical.

7 Acknowledgements

This project was supported, in parts, by Siemens Corporate Research contract 0305197992 and VACE Phase II contract 2004H840200000.

References

[1] Ahmed Elgammal, David Harwood, and Larry S. Davis, "Nonparametric background model for background subtraction," in *Proc. of 6th European Conference of Computer Vision*, 2000.

[2] W.E.L.Grimson and C.Stauffer, "Adaptive background mixture models for real-time tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1999.

[3] Antoine Monnet, Anurag Mittal, Nikos Paragios, and Visanathan Ramesh, "Background modeling and subtraction of dynamic scenes," in *International Conference on Computer Vision, Nice, France*, 2003.

[4] Bastian Goldlucke and Marcus A. Magnor, "Joint 3d-reconstruction and background separation in multiple views using graph cuts," in *Proc. IEEE Computer Vision and Pattern Recognition, Madison, Wisconsin*, Jun 18-20 2003, p. 683.

[5] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *IEEE Workshop on Stereo and Multi-Baseline Vision, Kauai, Hawaii*, 2001.

[6] Vladimir Kolmogorov and Ramin Zabih, "Multi-camera scene reconstruction via graph cuts," in *ECCV (3)*, 2002, pp. 82–96.

[7] K. Konolige, "Small vision system: Hardware and implementation," 1997.

[8] Christopher Eveland, Kurt Konolige, and Robert C. Bolles, "Background modeling for segmentation of video-rate stereo sequences," in *Proc. IEEE Computer Vision and Pattern Recognition, Santa Barbara, CA*, Jun 1998, pp. 266–271.

[9] Takeo Kanade, A. Yoshida, K. Oda, H. Kano, and M. Tanaka, "A stereo machine for video-rate dense depth mapping and its new applications," in *Proc. IEEE Computer Vision and Pattern Recognition, San Francisco, CA*, Jun 18-20 1996.

[10] G. Gordon, T. Darrell, M. Harville, and J. Woodfi ll, "Background estimation and removal based on range and color," in *Proc. IEEE Computer Vision and Pattern Recognition, Fort Collins, Colorado*, 1999.

[11] Yuri A. Ivanov, Aaron F. Bobick, and John Liu, "Fast lighting independent background subtraction," *International Journal of Computer Vision*, vol. 37, no. 2, pp. 199–207, 2000.

[12] Antonio Criminisi, Ian D. Reid, and Andrew Zisserman, "Single view metrology," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 123–148, 2000.

(a) (b) (c) (d)

Figure 6: (a) Two-camera initial detection. Red square marks a sampled pixel while white square marks its base, computed using weak perspective model. (b) Single camera detection in the second view used to constrain stereo searches. Red square marks the conjugate pixel and white line is the associated epipolar line. (c) Two-camera final detection; specular region is removed since it lies below the base point in (a). (d) Single camera detection in the reference view for comparison.

(a) (b) (c) (d)

Figure 7: (a) Two-camera initial detection. Here, the specular region was clustered as a separate bounding box. (b) No valid match could be found for the specular region. (c) The specular region successfully removed in the two-camera final detection. (d) Single camera detection in the reference view.

(a) (b) (c) (d)

Figure 8: (a) Some detected pixels remained near the top portion in the two-camera initial detection. (b) Single camera detection in the second view. The conjugate pixel was found at the top of the person. (c) Foreground filling gives a very good detection of the person even though he is very near the background wall. (d) Single camera detection in the reference view.

(a) (b) (c) (d) (e) (f)

Figure 9: Indoor scene. (a) Two-camera initial detection. Three sampled pixels are shown in red squares, while the blue squares are the bases. Noise near the shadow was eliminated in the final detection since it was below the base. (b) Single camera detection in the second view. Stereo matches are found for the three sampled pixels. (c) Two-camera final detection using only one of the sampled pixels; the lower boundary is not well-defined. Perspective model used here. (d) Comparison with the weak perspective model. The object was over-filled. (e) The three bases are used to form the lower boundary. A few more sampled pixels should fully recover the lower boundary. (f) Single camera detection in the reference view.