SELECTIVE SUBTRACTION WHEN THE SCENE CANNOT BE LEARNED

Adeel A. Bhutta[1] *Imran N. Junejo*[2] *Hassan Foroosh*[1]

[1]University of Central Florida, Orlando, FL, USA
[2]University of Sharjah, Sharjah, U.A.E

ABSTRACT

Background subtraction techniques model the background of the scene using the stationarity property and classify the scene into two classes of foreground and background. In doing so, most moving objects become foreground indiscriminately, except for perhaps some waving tree leaves, water ripples, or a water fountain, which are typically "learned" as part of the background using a large training set of video data. We introduce a novel concept of background as the objects *other than* the foreground, which may include moving objects in the scene that cannot be learned from a training set because they occur only irregularly and sporadically, e.g. a walking person. We propose a "selective subtraction" method as an alternative to standard background subtraction, and show that a reference plane in a scene viewed by two cameras can be used as the decision boundary between foreground and background. In our definition, the foreground may actually occur behind a moving object. Furthermore, the reference plane can be selected in a very flexible manner, using for example the actual moving objects in the scene, if needed. We present diverse set of examples to show that: (i) the technique performs better than standard background subtraction techniques without the need for training, camera calibration, disparity map estimation, or special camera configurations; (ii) it is potentially more powerful than standard methods because of its flexibility of making it possible to select in real-time what to filter out as background, regardless of whether the object is moving or not, or whether it is a rare event or a frequent one.

Index Terms— Background Subtraction, Scene Modeling, Dynamic Scenes

1. INTRODUCTION

Background subtraction is the fundamental step used in many applications including object detection, tracking, gesture and action recognition, activity recognition, and user interfaces. Background subtraction techniques traditionally use one or more views to classify the objects (or image pixels) as either foreground or background. However, standard methods have a rigid definition of what constitutes a background, which often leads to classifying almost all moving objects as foreground, except for small persisting motions that can be learned from a training set. This loss of 'intra-class separability' results in inability to model partial background or partial foreground and thus the notion of a background object being *in front* of a foreground object. If scene modeling is to be made more effective, the background subtraction techniques need to ensure that the statistical models can learn partial backgrounds and thus an intra-class taxonomy is preserved; which can prove very useful in many real world applications such as video surveillance and detection and tracking in crowds.

Existing background subtraction techniques can be classified into two main categories: techniques using monocular sequences and those using stereo sequences. Our method relies on two views but does not require configuring cameras rigidly as a stereo pair. Most of the existing literature focuses on aspects such as the statistical approach used to model the background, type of scene used (dynamic or static), the learning method applied to the training set, and the model used for the background or foreground. The background of a scene is generally defined as being *motionless* for static scenes (e.g., video conference) and *almost-motionless* for dynamic scenes (e.g., scenes which include changes such as illumination, shadows, waving tree leaves, water ripples, or fountains). Most single-view background subtraction techniques try to model the background (and the dynamic changes) either by modeling each pixel [1, 2] or different regions [3, 4] statistically, and then use those statistical models to detect the moving objects, known as foreground [5, 6]. This modeling requires large amount of training data for learning the statistical properties of the background. Alternatively, stereo-based techniques rely on estimating disparity maps by rectifying the views and using similarity measures in order to estimate the background [7, 8]. Such disparity maps are in practice difficult to estimate in real-time and very error prone. Also, these techniques require special camera setup and are computationally expensive. Furthermore, all background subtraction techniques classify moving objects as foreground indiscriminately. Consider a case when you have a crowded street with multiple objects moving across the camera in both directions. Any standard background subtraction technique will consider all of the moving objects as foreground thus will not be able to selectively distinguish which moving object should be kept as foreground and which ones discarded. What if you are

only interested in the first two objects closest to the camera, or only one object at the back, and all other objects are irrelevant. Thus, the *foreground-of-interest* is now the partial foreground while *background-of-interest* is a combination of traditional background and partial foreground. In this context, the standard definition of background is insufficient. Current background subtraction techniques fail to model such backgrounds.

Our technique has several novel contributions. Firstly, most background subtraction techniques require training or learning of the background model but cannot learn the partial background as defined above. We challenge the requirement of training and propose the use of a *reference plane* inducing a *base homography*, estimated using only two frames. This base homography can be used in the background subtraction of the scene when traditional technique fail, because they cannot classify an infrequently occurring moving object as background. Secondly, we propose to use the actual moving objects in the scene to estimate the base homography and show how a simple *walk* (or an object in motion) can be used to define a reference plane. Thirdly, standard background subtraction techniques fail to change the background model once it is learned. Only some minor dynamic changes are incorporated in the updating of the background model. In our proposed technique, the base homography can be modified using a different moving object or a plane in the scene in real time, and can be replaced altogether with a new base homography, thus providing flexibility in the background subtraction. Lastly, we avoid the explicit use of depth map and the requirement of rectifying two views for calculating depth as in stereo-based methods, and propose a solution based entirely on projective depth.

The rest of the paper is organized as follows. The theoretical formulation and the description of the proposed approach is presented in Section 2. Experimental results performed on real world sequences and brief discussion of results is presented in Section 3, followed by conclusion in Section 4.

2. SELECTIVE SUBTRACTION APPROACH

In this section, we first define selective subtraction and provide the theoretical formulation for implementing it.

2.1. Reference Plane π and Base Homography

Consider a sequence of images $\{\mathbf{I_t}\}^{t=1\ldots n}$, where multiple objects are moving across the scene as shown in Figure 1(a). A simple change detection algorithm can be used to detect the moving objects (or blobs) and their head and feet positions can be obtained by using the approach described in [9]. Let $\mathbf{P}_1$ and $\mathbf{P}_2$ be the two 3×4 camera projection matrices of two arbitrary cameras observing the scene. Since we do not require any calibration or a specific configuration, without loss of generality, we will model the two cameras as canonic cameras, i.e. $\mathbf{P}_1 = [\mathbf{I}, \mathbf{0}]$ and $\mathbf{P}_2 = [[\mathbf{e}']_\times \mathbf{F}, \mathbf{e}']$, where $\mathbf{F}$ is the fundamental matrix, $\mathbf{e}'$ is the epipole in the second camera view. Next, define the head and feet positions of a person viewed by these two cameras at a given instant in time as $\mathbf{p_1^t}$ (top), $\mathbf{p_1^b}$ (bottom) and $\mathbf{p_2^t}$ (top), $\mathbf{p_2^b}$ (bottom) points, respectively. These corresponding pair of points define a one parameter family of planes given by π_α, where α is a scalar parameter. The homography induced by this family of planes is then given by

$$\mathbf{H}_\alpha = [[\mathbf{p_2^t}]_\times \mathbf{p_2^b}]_\times \mathbf{F} + \alpha \mathbf{e}' \mathbf{p_2^b}^T [\mathbf{p_2^t}]_\times^T \quad (1)$$

Now, let $\mathbf{m}$ and $\mathbf{m}'$ be two corresponding points of a 3D point $\mathbf{M}$ viewed by the two cameras. The homography $\mathbf{H}_\alpha$ would map $\mathbf{m}$ from the left image to the right image as

$$\begin{aligned} \mathbf{H}_\alpha \mathbf{m} &= [[\mathbf{p_2^t}]_\times \mathbf{p_2^b}]_\times \mathbf{F}\mathbf{m} + \alpha \mathbf{e}' \mathbf{p_2^b}^T [\mathbf{p_2^t}]_\times^T \mathbf{m} \quad (2) \\ &= (1-\gamma)\mathbf{m}' + \gamma \mathbf{e}' + \beta \mathbf{e}' \quad (3) \end{aligned}$$

where $\beta = \frac{\alpha}{\mathbf{p_2^b}^T [\mathbf{p_2^t}]_\times^T \mathbf{m}}$, γ is a scalar parameter, and the last equation follows from the fact that the point $[[\mathbf{p_2^t}]_\times \mathbf{p_2^b}]_\times [\mathbf{e}']_\times \mathbf{m}'$ is on the epipolar line $[\mathbf{e}']_\times \mathbf{m}'$ and hence can be written as a linear combination of $\mathbf{e}'$ and $\mathbf{m}'$.

Therefore, by proper scaling of the last equation and rearranging, we can get

$$\tau = \frac{\mathbf{H}_\alpha \mathbf{m} - \mathbf{m}'}{\mathbf{e}' - \mathbf{m}'} \quad (4)$$

Here the scalar parameter τ may be interpreted as the projective depth of the point $\mathbf{M}$ from the plane π_α, because we can readily verify that if $\mathbf{M} \in \pi_\alpha$, then $\tau = 0$. Otherwise, τ will be either positive or negative depending on which side of the plane, $\mathbf{M}$ lies. We can determine τ from either x or y coordinates of the points $\mathbf{m}$, $\mathbf{m}'$, and $\mathbf{e}'$.

One last issue before we describe how we can use (4) for selective subtraction: The base homography $\mathbf{H}_\alpha$ as derived above is parameterized in terms of a scalar α. There are several ways we can determine α. One simple way is to use a pair of corresponding points between the two camera views to solve for α using (1). For instance, either the head or feet point correspondences of the person in the two cameras in a later frame can be used to determine α. In this way, a walking person would establish a reference plane.

2.2. Selective Subtraction

We use the reference plane as the decision boundary between foreground and background objects. *Any plane in the scene can be chosen as the reference plane and thus it gives us the flexibility of selectively keeping or subtracting the objects on either side of the plane*. For instance, if the reference plane chosen is the farthest plane in the scene then all moving objects fall in front of the reference plane and thus the approach can be used as a traditional background subtraction technique.

Fig. 1. Selective subtraction results for outdoor sequence: Input images (*first row*), Selective subtraction results with different *reference planes* (*second to fourth rows*), results from mixture of gaussian method (*last row*).

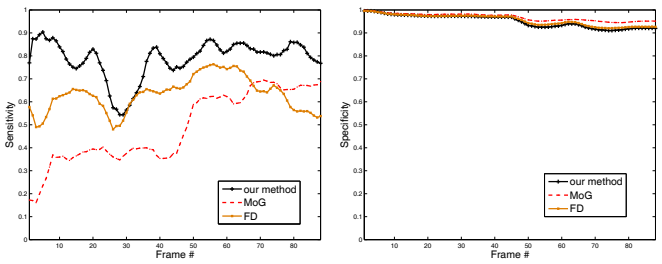
The projective depth (τ) for any moving object in the scene can be estimated and based on the sign of τ, the object can be classified as being on the foreground or the background. It is important to highlight that the proposed technique can also be used even when an object is fully or partially occluded (full occlusion can be detected as the object disappearing from the foreground).

3. RESULTS AND DISCUSSION

The algorithm was tested on a set of challenging sequences with multiple moving objects with significant occlusions and illumination changes. The comparative results with the mixture of gaussian method [1] have also been presented. The first sequence contains an outdoor scene with several moving objects along with shadows and dynamic motions including moving tree leaves. A simple frame difference algorithm with threshold along with connected component analysis was used for blob detection. The reference *walk* was selected as *reference plane* and *base homography* was estimated using head and feet positions in the first and the last frames. It should be highlighted that only four point correspondences are used to calculate the *base homography* and we do not require any additional training data.

We use Scale Invariant Feature Transform [10] to find point correspondences in the detected blobs and estimate the projective depth (τ) as described in Section 2.1. For each corresponding point, we calculate τ using (4) and use a majority voting scheme to classify the blob as foreground or background (i.e., as being on one side of the reference plane or the other). The results are shown in Figure 1 which show that the proposed algorithm can correctly separate the foreground from background.

One of the most unique aspects of our proposed technique is the flexibility it provides in selecting the *reference plane* of choice. Figure 1 shows how the foreground detection changes when different *reference planes* are selected for selective subtraction. First row of figure 1 shows input images as seen from the first view. Second row shows the results when the *reference plane* is the far wall and hence all moving objects are considered foreground as in traditional background subtraction technique. When the *reference plane* is changed to the farthest moving object (girl), the foreground changes accordingly as seen in third row. Fourth row shows the results when the selected *reference plane* is in the middle of pathway thus, detecting the objects in front as foreground. Notice that the girl walking to the left and the boy walking to the right are both on the other side of the *reference plane* and are detected as background. Furthermore, two boys walking to the left are correctly detected as foreground. Last row shows the results from mixture of gaussian method. Second test sequence containing the indoor scene with significant illumination changes was also used and the results are shown in Figure 3. First row in Figure 3 shows the input images from the first view. The objects found in front of the *reference plane* using selective subtraction are shown in second row and the results of mix-



(a) (b)

Fig. 2. Quantitative analysis of detection accuracy: (a) shows the sensitivity of the proposed algorithm (Average values: Ours 79%, [1] 49%, [11] 64%), (b) shows the specificity (Average values: Ours 95%, [1] 96%, [11] 95%).

ture of gaussian method are shown in the bottom row. The *reference plane* used in these results is the farthest wall in the scene. These results indicate that selective subtraction is effective and provides flexibility in selectively subtracting the objects of choice from the scene.

The qualitative analysis of the results shows that proposed technique performs better than mixture of gaussian [1] as shown in Figure 1, 3. We also performed the quantitative analysis of the pixel-level detection accuracy. The per frame detection rates are calculated in terms of sensitivity and specificity, as described in [4]. Figure 2 shows the sensitivity and specificity of the proposed technique as compared to [1] and [11]. Clearly, the detection accuracy in terms of sensitivity is consistently higher than [1] and [11] while specificity is comparable to both techniques.

One of the major advantages of the proposed technique is that it does not require any special camera setup as needed in other two-view background subtraction techniques. We also do not use the disparity map and thus the proposed algorithm is fast. It should be noted that we have not performed any post-processing, such as graph cuts [4] to improve the boundaries of foreground objects.

4. CONCLUSION

This work presents a number of fundamental innovations in the context of background subtraction. We present a novel concept of background as objects *other than* foreground which may include moving objects in the scene that cannot be learned from a training set because they occur only irregularly and sporadically. We propose a "Selective Subtraction" method as alternative to standard background subtraction, and show that a *reference plane* in a scene can be used as the decision boundary between foreground and background. Furthermore, the flexibility in selecting the *reference plane* using the actual moving object in the scene or an arbitrary plane in the scene, is truly unique to this method and is not available in existing background subtraction techniques. We present promising results on a challenging set of image sequences to show that the selective subtraction performs effectively and has applications in background subtraction, vehicle navigation, path anomaly detection, and detecting objects in crowds.

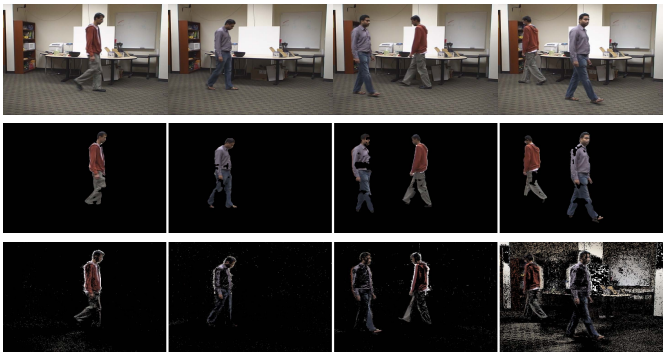


Fig. 3. Selective subtraction results for indoor sequence: Input images (*first row*), Selective subtraction results (*second row*), results from mixture of gaussian method (*last row*).

5. REFERENCES

[1] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *PAMI*, 2000.

[2] A. Elgammal, R. Duraiswami, D. Harwood, L. Davis, R. Duraiswami, and D. Harwood, "Background and foreground modeling using nonparametric kernel density for visual surveillance," in *Proc. of IEEE*, 2002.

[3] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh, "Background modeling and subtraction of dynamic scenes," in *ICCV*, 2003.

[4] Y. Sheikh and M. Shah, "Bayesian object detection in dynamic scenes," *CVPR*, 2005.

[5] T. Zhao, M. Aggarwal, R. Kumar, and H. Sawhney, "Real-time wide area multi-camera stereo tracking," in *CVPR*, 2005.

[6] R. Pless, J. Larson, Siebers S., and B. Westover, "Evaluation of local models of dynamic backgrounds," in *CVPR*, 2003.

[7] Y. Ivanov, A. Bobick, and J. liu, "Fast lighting independent background subtraction," in *ICCV Workshop on Video Surveillence*, 1998.

[8] S. Lim, A. Mittal, Davis L., and N. Paragios, "Fast illumination-invariant background subtraction using two views: Error. analysis, sensor placement and applications," in *CVPR*, 2005.

[9] T. Zhao F. Lv and R. Nevatia, "Self-calibration of a camera from video of a walking human," in *ICIP*, 2002.

[10] D.G. Lowe, "Object recognition from local scale-invariant features," in *ICCV*, 1999.

[11] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practices of background maintenance," in *ICCV*, 1999.