Active Background Modeling: Actors on a Stage

Raphael Sznitman  Henry Lin  Manaswi Gupta  Gregory Hager
Department of Computer Science  Johns Hopkins University  sznitman@jhu.edu

Abstract

Over the last two decades, background modeling techniques have focused on representing the general appearance of a background that is assumed to be predominantly static. However, there are many situations in which there are active, moving elements that are effectively part of the background. Examples include tools in manipulative tasks or work settings where a small, fixed set of people are moving about. Such situations are not well modeled by traditional methods.

In this paper, we present a background modeling approach, Actors on a Stage (AOS), that is able to accommodate both passive and active backgrounds. AOS is presented as a general, recursive estimation scheme for a background model. In this model, actors are a latent variable that is used to explain both occlusion of, and abrupt changes to, a background model. We demonstrate AOS in two different situations: a person writing on a blackboard, and a microretinal membrane peel. Additionally, we show how our method performs compared to traditional techniques in these settings, and on standard image sequences.

1. Introduction

Background modeling is an essential first step in many computer vision applications, where detection of objects that differ from the “normal” appearance is desired. Background modeling has seen use in many settings, including people [6, 2, 17, 13, 25] and vehicle [12, 5, 19] tracking, surveillance systems [16, 8] and behavior analysis [1, 14, 15].

Over the last two decades background modeling has addressed both static and dynamic backgrounds. However, in both cases, the background is viewed as a passive element that remains stationary (in a statistical sense) over time. Consequently, background models have lumped rapidly changing environmental elements (e.g. lighting, shadows, people, tools) as a generic statistical disturbance. Statistical models for backgrounds include Gaussians [25, 12], mixtures of Gaussians [21, 8], non-parametric distributions [4, 7] and Codebooks [11, 24]. All of these models have one common attribute: objects that appear or move in an image sequence are initially segmented as foreground, and if they remain static for relatively long periods of time they are incorporated into the background. As a consequence, normal, repetitive change (e.g. someone moving in and out of scene) causes continual, often unwanted, change detection.

We are interested in addressing a class of problems of this form. Namely, the background configuration may change over time and can be regarded as active. This is common in manipulative and interactive tasks, such as in surgical settings (e.g. retinal microsurgery, laparoscopic surgery). In these scenarios, the background is a surface of interest (i.e. eye retina), that is constantly occluded and manipulated by surgical tools. However, the goal is to maintain an accurate background model of the retina in spite of occlusion, while rapidly accounting for expected changes due to surgical manipulations.

In this paper, we introduce the notion of actors in a background model. Actors are active elements that are expected to appear in scenes, and which can be used to explain structural changes in the background. We refer to our background modeling method as Actors on a Stage (AOS). We formulate AOS as a time series estimation problem, where actors (tools, people, lighting, etc.) are a latent variable that can be used to explain large, abrupt changes to the background. We demonstrate that this framework, using established methods for object detection and relatively simple updating rules, outperforms most common background modeling techniques. We also show that, by using more expressive background models, AOS does not require an extensive training phase of the background model. Given a single initial observation of the background, our method sequentially improves its background estimate. This is particularly advantageous as training images may be extremely expensive or unavailable.

The rest of this article is organized as follows: Section 2 provides a detailed explanation of our model, where the general framework is presented in section 2.1 and an ap-
proximation is established in sections 2.2 and 2.3. Section 3 will demonstrate how our framework can be used in three scenarios: blackboard writing 3.1, micro-retinal peeling surgery 3.2 and in standard data sets 3.3. Some general comments are discussed in section 4 and we conclude with some final remarks in section 5.

2. Active Background Model

In this section, we will first formulate the general structure of our approach to background modeling, and then we describe the simplifications we use to achieve the results in this paper.

In what follows, we consider the background to be a random field, and we pose the inference problem on this field. Let \( \mathcal{U} = \{ u_1, u_2, \ldots, u_N \} \) be the indices of a set of observable image locations. We denote the background at time \( t \) as \( B^t \), where each location \( B_{u}^t \) will be considered a random variable. We infer \( B^t \) from an image sequence \( I^t = \{ I^0, \ldots, I^t \} \), where \( I^t_u \) is the observed value of an image at location \( u \) at time \( t \).

In order to model the effects of external actors, we introduce two additional families of random fields. First, background dynamics are often affected by some object (e.g. car, person, tool, \ldots). When an object changes the background, not only does the appearance of the background scene change, the background itself then changes. For this reason, we refer to these objects as actors and will denote the set of actors in a scene as \( A = \{ A_1, \ldots, A_n \} \). For subsequent notational simplicity, we will henceforth consider all of the background occlusion relationships to be represented in terms of a single “accumulator” \( A = \vee_{i=1}^{n} A_i \). We thus write \( A^t_u \) to denote the binary random variable representing whether some actor is occluding pixel location \( u \) at time \( t \).

Second, lighting and shadows change the appearance of the background but do not change the background itself. To this end, we associate fields \( L = L_1, L_2, \ldots, L_m \) with lighting effects. For example, in retinal surgery these fields would describe the lighting effects due to an endoscope (producing a local spotlight-like effect) and the effect of tool shadows. Note that the former can be described using a much smaller set of variables (e.g. the direction of the light source), in which case the latter is functionally dependent on an actor (the tool) and the light source direction. As before, for simplicity, we will consider a single random field \( L \) describing the cumulative effects of all lighting factors. For concreteness, \( L^t_u \) is a pair of random variables representing change in brightness and contrast at location \( u \) at time \( t \).

In what follows, let \( S^t = ( B^t, A^t, L^t ) \).

2.1. General Framework

We will cast the background inference problem as a Bayesian sequential estimation. That is, given an image sequence \( I^t = \{ I^0, \ldots, I^t \} \), infer \( S^t \). This can then be formulated as a Bayesian filtering problem \([9, 22]\) as

\[
P(S^t) = P(S^t | I^t)
= \int P(S^t | I^t, S^{t-1}) P(S^{t-1}) dS^{t-1}
\]

where (2) follows from (1) by an application of Bayes theorem, the assumption of Markov dynamics in the background model, and the assumption that \( P(S^t) \) is a sufficient statistic for \( I^t \).

We note that many of the extant approaches to background modeling can be viewed as a special case of this model. For example, if each background location is independent of every other, the dynamics of background change are linear, and background change and observations are contaminated with Gaussian noise, the optimal estimation scheme is the Kalman filter \([22]\). Nonlinear or non-Gaussian models can be estimated with approximate sequential Monte Carlo methods \([9]\).

Many approaches assume that a series of images is observed prior to estimation, e.g. \([18]\). For example, background models as mixtures of Gaussians \([21]\) or Codebooks \([11]\) can be viewed as batch generate or discriminative estimation problems, respectively. If we assume that image locations are not independent, then background modeling becomes a question of inference in a Markov random field.

We now describe the specific instantiation of this model used in this paper and derive the estimation techniques used.

2.2. Models

The dynamical model involves three components: the evolution of the underlying background, the evolution of the lighting model, and the evolution of the foreground actors. In the current implementation, we assume that the foreground object can be parameterized as a small set of parameters \( \alpha \) which correspond to the location and configuration of the object. Thus, we can write \( A(\alpha) \). Further, for the purposes of this paper we will neglect state history and uncertainty in \( \alpha \), and solve for a maximum likelihood estimate of \( \alpha \) using a detector. Details of how this estimate is computed vary by application and are given in section 3. Likewise, our lighting model will also be parametric, and will consist of a single global brightness change, \( \beta \). It will also be estimated independently on each frame using a maximum likelihood estimator as further described below.

Our nominal background model will be stationary with Gaussian noise. However, we will account for foreground occlusion and allowable scene change. To do so, we intro-
duce an intermediate random field $C$ where

$$ C^t_u = \text{dilate}(A^t, R) - A^t $$

(3)

where $R$ is a circular structuring element of radius $r$.

With this, we can write

$$ P(B^t_u|B^{t-1}_u) = (1 - \eta)N(B^{t-1}, \sigma^2) + \eta U(0, 255) $$

$$ \eta = \lambda_1 C^t_u, \quad 0 < \lambda_1 \leq 1 $$

(4)

where $U(\cdot)$ denotes the uniform distribution, $N(\cdot)$ denotes the normal distribution, and $\lambda_1$ is a design parameter governing the prior probability of background change by an agent.

The observation model is also modeled as a contaminated Gaussian depending on the location of all actors, the lighting model and the background. We formalize the observation model as

$$ P(I^t_u|B^t_u, A^t_u) = (1 - \gamma)N(B^t + \beta, \sigma^2) + \gamma U(0, 255) $$

$$ \gamma = \lambda_2 F^t_u, \quad 0 < \lambda_2 \leq 1. $$

(5)

where $F^t_u$ is a random field similar to $C^t_u$, but which denotes the regions which are possibly affected by the tool actions.

### 2.3. Model Updating

Based on the definitions above, there are three cases when estimating the updated background model:

$F^t_u = C^t_u = 0$: This is an estimate of the nominal background in areas beyond the change radius of any agent. In this case, the optimal solution to (2) under the models in (4) and (5) is a Kalman filter modulo the brightness value $\beta$, which functions as a latent variable. We solve for $\beta$ as the average difference

$$ \beta = \frac{\sum_u (I^t_u - B^{t-1}_u)(1 - F^t_u)}{\sum_u (1 - F^t_u)} $$

(6)

We then implement updating as a steady-state Kalman filter which yields a simple weighted averaging rule:

$$ B^t_u = (1 - \alpha)B^{t-1}_u + \alpha(I^t_u - \beta) $$

(7)

where $\alpha$ is the steady-state Kalman gain.

$F^t_u = 1$: We set $\lambda_2 = 1$. In this case, by (5), observations provide no information and

$$ B^t_u = B^{t-1}_u $$

$C^t_u = 1$: In this case, we set $\lambda_1$ to a small value based on the expected probability of a change in the background near an object. As a result, the posterior distribution of (2) will be multi-modal in cases where the observed value and the prior background value differ greatly. We compute the minimum distance $\tau$ such that the distribution develops two distinct modes. Define

$$ R^t_u = \begin{cases} 
1, & \text{if } |I^t_u - \beta - B^{t-1}_u| > \tau \\
\alpha, & \text{otherwise}
\end{cases} $$

where $\alpha$ and $\beta$ take the same values as in (7)

The update equation is then

$$ B^t_u = R^t_u (I^t_u - \beta) + (1 - R^t_u)B^{t-1}_u $$

These updating rules are intuitive as they tackle the possible scenarios of the observations. In the first case, the image may not have changed, or may have slightly changed due to minor illuminations. The second case accounts for when change has been observed and an actor has been recognized at the same location. For this reason, the old background is kept and a mask is applied in order to account for the actor. The third case is where the background has changed, due to an observed $A_t$. In this case, the new background is made of the new image information.

#### 2.3.1 Background Initialization

In general, two types of images may be available at time $t = 0$: Either only the background is visible, or actors are present in the image. In order to solve both cases, we initialize $B^0 = I^0$ and then remove pixels which are occluded by actors in the image. In the context of the model described thus far, layer $A$ is subtracted from $B^0$. The justification for this is simple: if an object is occluding the background, then no assumption may be made on the structure being occluded. As time elapses, filling-in of unseen regions will be done by the background updating procedure (section 2.3).

### 3. Applications

In this section, we present two examples of active backgrounds, and show how our proposed method maintains an accurate estimate of the background. The first involves a person drawing on a blackboard, and the second is a segment from a live micro-retinal peeling surgery. In both cases, we show how AOS and two other state-of-the-art methods for background modeling perform compared to manually segmented sequences. We also highlight algorithm performance differences on representative frames. Finally, we show how AOS performs on a standard image sequence dataset [10].

#### 3.1. Blackboard Writing

The first example uses an image sequence captured with a standard video camera ¹ of a blackboard in a classroom.

---

¹Canon Powershot SD630, 640 × 480 pixel RGB video
The sequence begins with no foreground objects in the field of view. After 50 frames, an individual approaches the blackboard and erases the scribble on the board. He then proceeds to draw a house. In this scenario, the blackboard is considered an active background because the individual is actively modifying it. Specifically, the erased scribble and drawn house are consequences of actions that change the existing background and thus should not be considered part of the foreground.

To accurately estimate the current background, we must specify the actors in the scene and their associated behaviors. We set $|A| = 3$, such that $A_1 = \text{person}$, $A_2 = \text{arm}$ and $A_3 = \text{hand}$. The division of an individual into three actors allows for better detection rates and modeling of the behavior of actors.

To detect $A_1$, we let $f_1$ be a Boosting classifier [23] and is trained according to [20]. Classifier $f_1$ returns a bounding box when a person is detected at a given location in an image. We let $r_1 = 1$ to indicate that $C_1$ cannot interact with the background.

$A_2$ is similar in structure to $A_1$. That is, $f_2$ is similarly trained as $f_1$. $r_2$ is similar to that of $r_1$.

$A_3$ is the most important actor in this scenario, since it is the only one that can directly alter the background structure. Boosting classifiers are ill-suited for $A_3$ because the hand is very small and generally of poor image resolution. We instead use normalized cross-correlation to match templates of hands to regions in the image. Since $A_3$ directly changes the background, $r_3 = 2 \cdot \text{size(hand)}$, which allows the background near the hand to be updated.

Experiments and Results

We tested AOS on the blackboard image sequence, along with two standard methods: Mixture of Gaussians (MOG) [21] and Codebooks (CB) [11]. We let both of these methods train on 50 initial frames, where no object is in the field of view. As a basis of comparison, we manually segmented the foreground of a representative segment (200 continuous frames) from the sequence. For each algorithm, we computed the average error per frame, $E_{\text{avg}}$, between the foreground/background segmentation result and the manually segmented foreground. We also computed the average background error, $B_{\text{avg}}$, for each method. Specifically, $B_{\text{avg}}$ is an indication of how well the background specifically is being estimated.

In Table 1, we show how all three methods perform on these 200 frames. AOS is $\approx 4$ percentage points more accurate in estimating the background and the foreground in the image sequence. More importantly, the $B_{\text{avg}}$ for AOS is significantly lower than for MOG and CB in the blackboard segment: 8% versus 14% and 15%.

Figure 1 shows how each algorithm performed on a small set of images from the sequence. In the first column, the person has just approached the board and has begun to erase. Here, AOS detects the individual better because it does not require a training phase. Once the initial writing is erased, both the MOG and CB methods have difficulty updating the active background. In the MOG case, the person has stayed in place too long and has become a part of the background. For CB, the erased writing remains a part of the foreground (Similar effects are seen in column 3). The CB also incorrectly identifies the new drawing as part of the foreground (column 4). However, in AOS, the writing that is erased and then drawn is considered part of the active background.

3.2. Micro-Retinal Peeling

A second application of active backgrounds is in the context of surgical manipulative tasks. In this example, the im-
Figure 2. Segmentation results of MOG, CB and AOS on four representative frames of the micro-retinal peeling example. The manual segmentation is shown in row 2. (1) Positioning of gripper above membrane: Foreground detection by MOG and CB is not optimal during early training period. (2) Initial peeling of the membrane: Parts of the gripper have been fused into the background in MOG. The partially peeled membrane is incorrectly detected as foreground by CB. (3,4) Peeling is in progress: The peeled membrane is incorrectly detected as foreground by MOG and CB.

The image sequence consists of a surgeon’s view during a live retinal membrane peel, where the surgeon attempts to remove a thin membrane on the eye retina. The retina can be viewed as the object of interest, and thus requires a consistently accurate estimate. In the image sequence, the surgeon slowly grips a section of the membrane, peels it off and reveals the bare retina. The image sequence begins with the surgical tool already in the field of view.

The actors in this image sequence consist of $A_1 = \text{closed\_gripper}$, $A_2 = \text{open\_gripper}$ and $A_3 = \text{shadows}$. We make the distinction between an open and closed gripper because the retina can only change when the gripper is closed (gripping membrane). $A_1$ and $A_2$ are similarly detected, but $C_1$ and $C_2$ are different. That is, since only $A_1$ can affect the background, $r_1 = 2 \cdot \text{size(tool)}$ and $r_2 = 1$. $f_1$ and $f_2$ are detected in a similar way as in [3], templates of the tool are extracted and then matched in the image.

Detecting $A_3$ is done as follows: An initial computation of the gradient image is performed on both $B^{t-1}$ and $I^t$. For each pixel in each image, a histogram of oriented edges is then computed around a small window (7x7 pixels in the experiments). At a particular location, if two histograms (one from $B^{t-1}$ and the other from $I^t$) differed enough, the pixel is labeled as shadow. We also set $r_3 = 1$, such that shadows do not change the retina.

**Experiments and Results**

For the micro-retinal peeling, MOG and CB were trained on 100 initial frames with the tool in the field of view. 180 continuous frames were manually segmented as a basis for comparison. The Retina column in Table 1 show the average errors, $E_{\text{avg}}$ and $B_{\text{avg}}$, of the three methods for the frame sequence. The CB method has the most difficulty in correctly identifying the active background: ≈ 8 percentage points worse than the other two. AOS performed slightly better than the MOG method for both measures.

However, the first column of Figure 2 shows that both MOG and CB have difficulty detecting the gripper in the early stages of the sequence because the training is not complete. Since AOS did not require a training phase, the gripper is accurately detected and segmented. The next three columns show that even as the membrane has been peeled away, the underlying retina is still detected as background. Our method recognizes the initial peel (column 2) but quickly updates the background accordingly (columns 3 and 4). Generally, from Figure 2 and Table 1 our method provides a good model for retinal modeling.

**3.3. Wallflower Dataset**

In this section, we test our algorithm on the dataset presented in [10]. This dataset consists of 7 general conditions in which various actions are taking place: A moved object, smooth lighting change, global lighting change, camouflage, aperture problem, dynamic background and bootstrapping. We show how our method performs in each of these cases. In particular, one condition involves moving an object in a scene, and then treating this newly moved object as part of the background. A more complete analysis of this scenario is done, as this can be considered an active background.

For each of the image sequences, $|A| = 2$, such that $A_1 = \text{person}$ and $A_2 = \text{shadows}$. $f_1$ and $C_1$ are designed in a similar fashion as the $A_1$ in section 3.2. The only parameter that differs in each image sequence is $r_1$ because the effect of $A_1$ is dependent on the respective scenario. We detect $A_2$ the same way as the $A_3$ in section 3.2.

**Experiments and Results**

We ran AOS, MOG and CB on all 7 scenarios of the dataset. The image sequences consist of 160x120-pixel colored images. Parameters of MOG and CB were set for the best performance possible for each sequence, and were trained on the same training images as in [10].

Figure 3 shows how AOS performed on each image sequence. The foreground of the scene is displayed for one
Figure 3. Results on the Wallflower dataset. Top: Original images from each sequence. Middle: True foreground for an image in the sequence. Bottom: Computed foreground for each sequence on this particular frame. Note, that this is the test frame provided in [10].

particular image in each sequence (Last Row). Note that the test images used are those in [10]. Similarly, results for MOG and other algorithms are also available in this paper.

In general, we note that our results are as good as, and in some cases better than, those from previous methods on this dataset. In the case of the Waving Tree, the tree remains part of the foreground. This is because our implementation does not allow for multiple hypotheses of the background.

The scenario involving a moved object is a particularly interesting one in the context of active backgrounds. Figure 4 shows AOS, MOG and CB performed at various stages of the image sequence. The problem with MOG is that as time elapsed, motionless actors become part of the background. This is shown in column 1 of Figure 4 where the person is talking on the telephone. On the other hand, AOS correctly handles motionless actors. This is in sharp contrast to MOG or other adaptive methods which will include changed portions of the background proportionally to the learning rate $\alpha$. Notice that in both the MOG and CB, once the person gets up from the chair, both the telephone and chair (part) are treated as foreground.

4. General Discussion

From the model and the results presented in the sections above three important issues are important to point out.

First, in past methods, the variance in an image sequence was handled at the image level, usually in a training phase. In AOS, image variation is accounted for at the actor level. By learning what actors appear as, and detecting them through the frames, we remove the need for a training phase on the image sequence. Consequently, estimates of the background are more accurate in scenarios where objects are in the field of view all the time (section 3.2 and 3.3).

Second, background estimates are highly dependent on the accuracy of actor detection. In our examples, detection of actors is characterized by a bounding box. Consequently, regions of the bounding box that are not part of the actor would be incorrectly characterized as such. This inaccuracy can be seen as slight artifacts in our results (e.g. section 3.3, scenario 5 and 7). Ideally, a detection method that simultaneously locates and segments the actor would resolve these inaccuracies.

Finally, modeling dynamic backgrounds (e.g. ocean view, waving trees, rain) was not implemented in the examples above (see section 3.3, scenario 3 and 4). We believe the handling of multiple instances of the background is not an intrinsic problem with the model and is simply an
implementation issue.

5. Conclusion

We have demonstrated in this paper that active backgrounds can be accurately estimated by the detection of expected actors in the scene. This provides us a simple procedure to update the background model according to the actors’ behavior. We presented two applications involving active backgrounds. In the case of a person writing on a blackboard, we showed that we were able to accurately and quickly update the background, even if foreground objects do not change positions over extended periods of time. We also demonstrated the performance of our method, AOS, in real retinal membrane peeling image sequences, where a complex manipulation takes place. Finally, we compared the performance of AOS to a standard dataset and to state-of-the-art background modeling techniques, MOG and CB.

Acknowledgement

This work was supported by NIH BRP 1R01EB007969-01A1 and NSF Grant 0534359. The interpretations of the findings presented here lie solely with the authors.

References


