FOREGROUND SEGMENTATION IN SURVEILLANCE SCENES CONTAINING A DOOR

Andrew Miller and Mubarak Shah

Computer Vision Lab at University of Central Florida
Orlando, Florida 32816
amiller@cs.ucf.edu / shah@cs.ucf.edu

ABSTRACT

We propose a new method for performing accurate background subtraction in scenes with a door, like a building entrance or a hallway. This kind of scene is common in surveillance applications, yet the sporadic motion of a door causes problems for existing systems that falsely report the door as foreground. Our method models the scene's appearance by storing a set of gaussian pixel distributions corresponding to a discrete sample of the door's range of motion. All of the pixels in the image are dependent on the position of the door, so we use the joint probability for all of them to estimate the maximum-likelihood position of the door. We then perform background subtraction using the specific appearance model indexed by our estimated position. We show that our algorithm accurately segments the foreground region in several actual indoor and outdoor surveillance settings.

1. INTRODUCTION

Some of the most popular surveillance settings are centered around a door, such as a building entrance or an office hallway. Unfortunately, doors present difficult problems to most surveillance systems since they tend to violate basic assumptions about the nature of a background. Doors move relatively infrequently compared to camera noise or jittering clutter objects, and their appearance can change dynamically as they sweep out different angles, possibly reflecting a light source at the camera. When a person dynamically occludes a moving door, both the person and the door will be lumped together as a single foreground object.

Therefore we propose a solution that exploits a different property of the background to distinguish it from the foreground. The position of the door is a parameter of the entire scene, so we can use the *joint evidence* from every pixel in the region to determine the position of the door. Once we recover the position of the door, we unambiguously know appearance of the background and can perform background-subtraction just as easily as if the background were static.

2. RELATED WORK

The goal of a background-subtraction approach is distinguish the background from the foreground by utilizing some discriminating characteristic between them. Most systems rely on the assumption that the foreground will usually appear different from the foreground and treat each pixel as an independent sensor.

Pfinder was the original work in statistical background modeling, using a single RGB gaussian for each pixel in the background [1]. Pixels whose values differ substantially from the background model are marked as foreground. This system is simple and works effectively in scenes with a static background

Stauffer and Grimson [2] have developed a Mixture of Gaussians (MoG) background model which has become very popular because of its flexibility and stability in complicated scenes. It accounts for dynamic environments by allowing several surfaces to be modeled for each pixel. The task of determining whether a surface is background or foreground is accomplished by exploiting two heuristics: the background surfaces will appear more frequently than the foreground and will have narrower color distributions. This method works particularly well in scenes with high frequency multi-modal color distributions, such as outdoor scenes with fluttering leaves or camera jitter. However, a door is opened relatively infrequently compared to these situations, so it will still be classified as foreground even if it is recognized as a mode. Also, a door has several surfaces that rapidly change appearance through its range of motion, further confusing the algorithm. Power and Schoonees [3] provided a tutorial elaborating on the theoretical basis of MoG and suggesting parameter values.

Rittscher and Blake[4] used information about the temporal consistency of surfaces to help disambiguate between the overlapping appearance distributions of background, foreground, and shadows in a highway monitoring system by developing a probabilistic Markov model for transitions between these states. Although this method is notable for its use of additional information besides color to discriminate between labels, it is only effective in a scenario with a consistent temporal behavior, such as a highway, which has a defined speed limit and stable traffic flow.

More recently, systems have challenged the notion of pixel independence and incorporated information from spatially adjacent regions. Sheikh and Shah[5] adaptively construct a nonparametric distribution estimation that includes evidence from neighboring pixels. This method works particularly well in scenes with high frequency clutter motion, such as ripples on a body of water. A door is still problematic since it may move faster than the adaptive algorithm can propagate information spatially from neighbor to neighbor. This algorithm also relies on the higher frequency of background appearances to distinguish between background and foreground, so it will misclassify an infrequently opening door even if it anticipates the moving surfaces.

The topic of doors has received explicit attention from the mobile robotics community. Since doors are entryways to new rooms and areas, they are of interest to path-planning and map-making robots. Stoeter[6] developed an algorithm to automatically detect doorways in a scene by observing vertical edge features from camera images fused with laser range information, while Anguelov [7] used a generative probabilistic model of a hallway to generate a maximum-likelihood map of the walls and doors from visual and range data.

Our method is related to the previous efforts in background subtraction because we use a probabilistic scene model to derive a maximum-likelihood segmentation. Our original contribution is the application of this approach to a new problem, scenes with a door, and the use of problem-specific constraints to improve the accuracy of our segmentation.

3. PROPOSED METHOD

3.1. Modeling the Scene

Our scene model is derived from four assumptions. First, each pixel is an observation of either the background or a foreground object. Second, the appearance of the background is static except for the movement of the door. Third, nothing is known about foreground objects, so they are assumed to have a uniform color distribution [3]. Fourth, the actual color observation of each pixel is independently sampled from a normal distribution centered around the true color of the observed surface, with an empirically known covariance matrix. We assume that each of the RGB color channels are independent and have an identical variance, [2], and also that this variance is the same for every pixel. This becomes a system parameter, σ^2, that controls the tolerance for camera noise, compression artifacts, and small illumination changes.

We describe the foreground/background state of each pixel with the variable $k_{x,y} \in \{FG, BG\}$. The ultimate goal of background subtraction is to recover the value of this variable for each pixel. We conventionally assume that this random variable has no correlation between pixels or from one frame to the next, and that the prior probability of each state is empirically known through the parameter T [2, 3]:

$$P(k_{x,y} = BG) = T, \tag{1}$$

$$P(k_{x,y} = FG) = 1 - T. \tag{2}$$

This parameter effectively controls the *gain* of the system, or how eagerly the system tries to classify a pixel as foreground.

The position of a door can be described by a continuous variable $d \in [0, 1]$ ranging from fully closed to fully open. For practical purposes, we approximate this interval with a discrete set of states spanning the full range of the door's motion. Since the door is the only source of dynamic behavior in the background, the appearance model of the background consists of a mean color for each pixel and door position, $\mu_{x,y,d}$. Thus the color distribution for a given pixel, assuming the door position is known, and assuming that the pixel is in the background, is given as:

$$p(I_{x,y}|d, k_{x,y} = BG) = \frac{1}{(2\pi\sigma^2)^{\frac{3}{2}}} e^{\frac{-1}{2\sigma}(I_{x,y}-\mu_{x,y,d})^T(I_{x,y}-\mu_{x,y,d})}. \tag{3}$$

The model of the scene is generated from a short training sequence in which there are no foreground objects, and in which the door sweeps out a range of motion from fully closed to fully open. Four frames from such a sequence are shown in Figure 1. This sequence serves two purposes. First, we use this sequence to discretize the position of the door by splitting the interval $[0, 1]$ evenly among the frames. If there are ten frames in the sequence, then the first frame corresponds to $d = 0.0$, the second frame to $d = 0.1$, etc. Second, we directly use the color values from each frame as the mean of the color distribution for each pixel at each door position. Note that in this model, the variable d does not correspond directly to the actual angle of the door, but to a nonlinear mapping of the door angle that is affected by the framerate of the camera and the speed at which the door is opening. This is fine since we are not interested in the telemetry of the door, but only in its appearance at each position.

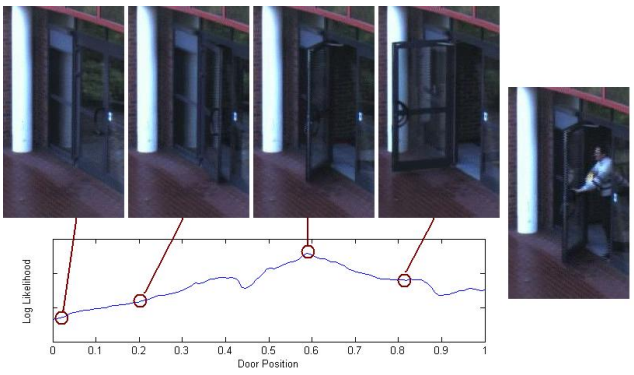

Fig. 1. The top four frames are samples from a training sequence where the door moves from fully closed to fully open. In the test frame to the right, the position of the door is most similar to the third sample frame. The log-likelihood graph indicates the most likely position of the door. The arrows indicate the correspondence between the graph and the sample frames.

We can marginalize the distribution over the variable k by summing the gaussian background distribution and the uniform foreground distribution, weighted according to their prior probabilities indicated by T:

$$p(I_{x,y}|d) = T \cdot p(I_{x,y}|d, k_{x,y} = BG) + (1 - T). \tag{4}$$

3.2. Determining the Door's Position

The first step of our method is to estimate the position of the door by considering the pixel values in each frame as evidence only of the door position d and not of the state variable k. Since the distribution in (4) provides a forward model of the pixel observations given the position of the door, we can use Bayes's rule to work backwards and determine the most likely door position given the evidence from the image and the uniform prior distribution of door positions. The value of each pixel is independently sampled from the corresponding distribution according to d; the pixel values are *conditionally independent* given the position of the door. Thus the joint distribution of the entire image I is given by the product of the distributions for each pixel:

$$p(I|d) = \prod_{x,y} p(I_{x,y}|d). \tag{5}$$

Using Bayes's rule, the posterior likelihood of the door's position given the evidence of the current frame is:

$$p(d|I) = p(d)\frac{p(I|d)}{p(I)}. \tag{6}$$

Since $p(I)$ is the same for every door position and $p(d)$ is assumed to be uniform, the most likely position is simply the value of d which maximizes $p(I|d)$. We have constrained d to a discrete set of values, so we can calculate the likelihood for each door position and choose the best for our estimate. The product of probability densities over a large number of pixels will tend to grow immeasurably small, so it is more effective to instead calculate the *log-likelihood*. An example of the log-likelihood plot for a test frame is shown in Figure 1.

Note that the effect of assuming a uniform likelihood for d in a Bayesian system, rather than something more powerful such as a Markov model as in [4], is to allow evidence alone to dominate the decision instead of influencing the decision with prior expectations. We assert that in this application the evidence is sufficient to determine the correct door position. As an alternative justification, consider that door-opening events are relatively infrequent, so a probabilistic model will tend to believe the door is always closed. Since the moments when the door is moving are more interesting than when it is still, our imposition of relative importance would effectively cancel out the relative frequency of door positions.

3.3. Background Subtraction

Once we have an estimate for the position of the door, $\hat{d}$, we know the appearance of the background. We can perform background subtraction by choosing the most closely matching frame in the training sequence and using it as a static background model as in [1]. A pixel is marked as foreground if it is more likely to have been sampled from the foreground distribution than the background distribution.

$$\hat{k}_{x,y} = \begin{cases} BG, & T \cdot p(I_{x,y}|\hat{d}, k_{x,y} = BG) > 1 - T \\ FG, & otherwise \end{cases} \quad (7)$$

The disadvantage of using a discrete approximation for a continuous variable is that the door position in a test frame will not *exactly* match a position in the training sequence. A way to compensate for this is to perform background subtraction by comparing each pixel in the test frame to the background distributions of several of its *spatial neighbors* rather than just to its own corresponding distribution. We use the eight adjacent neighbors as well as the current location, and weight each distribution equally: $p_{new}(I_{x,y}|\hat{d}) = \frac{1}{9}\sum_{i=-1}^{1}\sum_{j=-1}^{1} p(I_{x+i,y+j}|\hat{d})$. This is analogous to convolving an image with a blurring filter, only instead of summing each *pixel value* in the sliding window, we sum each *distribution*. This permits each pixel some spatial uncertainty to account for the subtle motion of the door that cannot be captured by the discrete model.

4. RESULTS AND ANALYSIS

To demonstrate that our algorithm effectively removes the door from the foreground segmentation, we collected a diverse dataset of typical indoor and outdoor surveillance settings, and compared the results of our algorithm to the MoG method[2]. We used the same parameters for both algorithms where applicable: initial variance σ^2 (either 20 or 30 out of 256, depending on the scene) and background probability T (0.9). For the remaining parameters of MoG, we used the default values as suggested in [3]. We created bounding boxes around the detected foreground objects by performing a morphological opening operation and 8-connected components.

In Figure 2 we show five frames from a video sequence of a person entering a building from an outdoor entrance. The first frame shows the background with the door closed, and no visible foreground objects. In the second frame, the door has begun to open but the person is not yet visible. Since it is nighttime, the glass panels in the door are reflecting the building interior rather than permitting a view outside. In the final three frames the person and door are both marked as foreground by MoG, but our algorithm the marks only the person as foreground.

Figure 3 shows more results from our own collected video. Since the elevator door in *g-j* consists of a uniform surface, MoG successfully learns to identify both the metal door and the wooden interior as background. However, since it considers each pixel independently, the background model is more ambiguous, resulting in an inaccurate foreground segmentation in frame *j*. Our method produces a more accurate segmentation since the joint evidence from all the pixels gives us a more specific background model.

Frame *j* suggests that this method could also be used to identify people getting in or out of a car. A model of the specific car appearance must be known ahead of time and the car must be precisely located in the scene. Since the background model in MoG does not consider the motion of the door, it adds the door to the detected foreground object while our method successfully segments just the person. Similarly, the joint-appearance-evidence of a continuous state variable may possibly be integrated into an object detection and recognition system.

Figure 4 shows results taken from actual surveillance cameras. In these scenes a corner of the door is often falsely detected as a single foreground object by MoG. If the door occludes the person, as in *p* and *q*, then the resulting foreground may be split or have reduced area when the our method removes the door door from the foreground segmentation.

The dataset from for *l-m* was provided by ETISEO, an evaluation funded by the French government for surveillance applications.

5. CONCLUSION AND FUTURE WORK

The use of joint evidence from all the pixels in the region is an effective way to determine the position of the door, and our method successfully removes the door from the foreground segmentation.

A weakness of this algorithm is that it depends on an accurate training sequence with no foreground objects, which may be difficult to obtain. It would be an improvement if the algorithm could 'bootstrap' itself by developing and updating a background model online, even in the presence of foreground objects. Also, the assumption of an static background with a single door may not hold in some scenes. If we can automatically detect the doors as in [6, 7], then the system could cope with multiple doors and other sources of dynamic behavior by fusing our method with a more robust general-purpose background subtraction.

This research was funded in part by the U.S. Government VACE program.

6. REFERENCES

[1] C. R. Wren et al, "Pfinder: real-time tracking of the human body," *in PAMI*, July 1997.

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

[3] W. P. Power and J. A. Schoonees, "Understanding background mixture models for foreground segmentation," in *Proceedings Image and Vision Computing New Zealand*, 2002.

[4] J. Rittscher et al, "A probabilistic background model for tracking," in *in Proceedings of ECCV*, June 2000.

[5] Y. Sheikh and M. Shah, "Bayesian modeling of dynamic scenes for object detection," *PAMI*, October 2005.

[6] S. A. Stoeter et al, "Real-time door detection in cluttered environments," in *In IEEE International Symposium on Intelligent Control*, July 2000.

[7] D. Anguelov et al, "Detecting and modeling doors with mobile robots," in *Proceedings on ICRA*, 2004.

Fig. 2. Key frames from results on a video sequence. The first row of each block contains original frames, the second row shows the results of MoG background subtraction, and the third row shows the results of our algorithm. MoG falsely labels the door as part of a foreground object, while our algorithm accurately detects just the person.

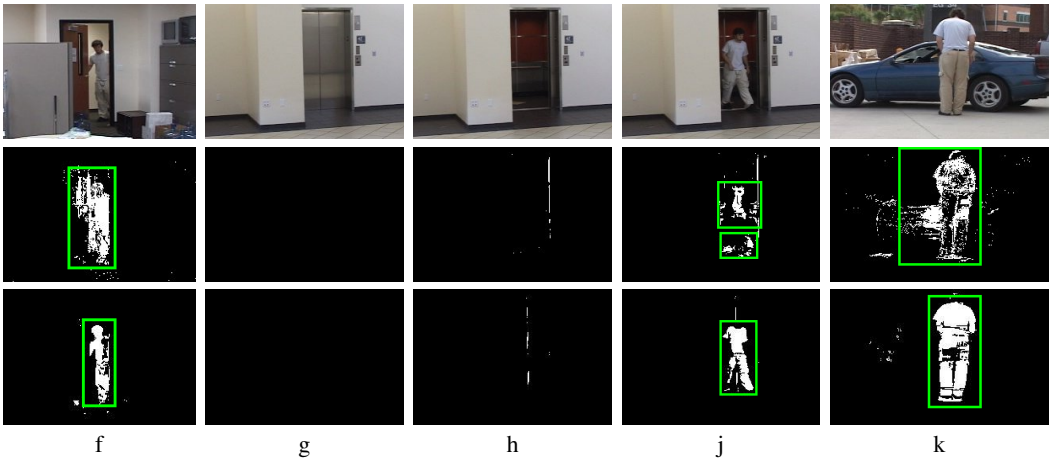

Fig. 3. In *f* our algorithm accurately removes the visible door from the detected foreground object. In *g* and *h*, MoG correctly identifies the elevator as background when the door is open and closed. However, since it treats each pixel independently, it does not have enough information to make an accurate foreground segmentation in *j*. In *k*, we show that our algorithm could also be used to segment a person getting in or out of a car.

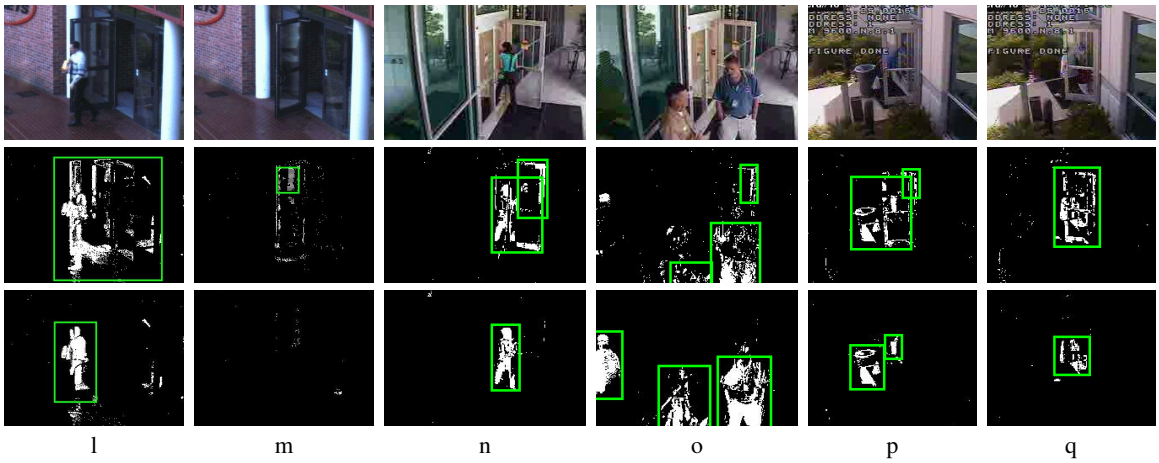

Fig. 4. Results from actual surveillance videos. If the door occludes the person, as in *j* and *k*, then the resulting foreground may be split or have reduced area when the door is removed from the foreground segmentation.