A BAYESIAN FRAMEWORK FOR GAUSSIAN MIXTURE BACKGROUND MODELING

Dar-Shyang Lee, Jonathan J. Hull, Berna Erol

Ricoh California Research Center
2882 Sand Hill Road, Menlo Park, CA 94025, U.S.A.

ABSTRACT

Background subtraction is an essential processing component for many video applications. However, its development has largely been application driven and done in ad hoc manners. In this paper, we provide a Bayesian formulation of background segmentation based on Gaussian mixture models. We show that the problem consists of two density estimation problems, one application independent one dependent, and a set of intuitive and theoretically optimal solutions can be derived. The proposed framework was tested on meeting and traffic videos and compared favorably over well-known algorithms.

1. INTRODUCTION

Background subtraction is an early and essential processing component for many video application systems. Its goal is to separate out the *foreground*, or more accurately the *region of interest*, from the rest of the video. Over the years, background models have evolved from a static, single Gaussian representation into more complex models such as adaptive Gaussian mixtures to handle real world conditions. Several researchers have applied Gaussian mixtures for background modeling in surveillance applications. However, the discussion often focused on overall system performance, and background segmentation was often applied in an ad hoc manner.

Many methodologies for background segmentation have been reported, almost one in every object tracking system. Typical solutions rely on a training period to collect statistics of the true background represented by mean and variance [10] or a pair of bounds [3]. Recently, automatic background estimation based on more complex models have been proposed [1,2,8]. We favor the Gaussian mixture model [8] over the non-parametric, kernel-based approach [1] for its efficiency and analytical form. However, the decision problem at the pixel level has received little formal treatment, largely due to the fact that it is never intended to be used alone. Consequently, its usage tends to be application driven and ad hoc.

Some formulations of this process have been made, however, usually in a very constrained context. For example, a framework based on explicit modeling of the illumination and surrounding noise was proposed [7]. The system is not adaptive to scene changes and requires a training period, leaving limited application context. A more general treatment of the problem was provided in [9] who outlined various properties a background modeler should possess and proposed a three level

processing framework. However, no framework at the pixel level processing was offered.

In this paper, we present a Bayesian formulation of the background segmentation problem at the pixel level based on Gaussian mixture modeling. Gaussian mixtures are not only more adequate for modeling real world data, its analytical form also fits well into a statistical framework. With an explicit expression of the underlying pixel distribution, the original background segmentation problem can be formulated as two independent density estimation problems. The first problem is to model the distribution of values observed at each pixel location with a Gaussian mixture, which is application independent. The second is classification of constituent Gaussians as foreground or background, which is inevitably domain specific. We also derive an explicit formulation of the background model and its representation, which has not been addressed in earlier works.

The rest of the paper is organized as follows. In Section 2 we describe the framework and provide a formulation for the background model. We describe our solution to the two density estimation problems in Section 3. The proposed framework is tested on real meeting and traffic videos and compared to the algorithm of [8]. The results are presented in Section 4, followed by conclusions.

2. BAYESIAN FRAMEWORK

2.1 Framework Overview

At the lowest level, video background segmentation is a binary classification problem: decide each pixel in the frame at time t as *foreground* or *background*. From a Bayesian perspective, this decision should be based on the posterior probability of the pixel being background $P(B|x)$ where x denotes the pixel observed in the frame at time t and B denotes the background class. Without giving a precise definition of *foreground* and *background*, which is most likely application dependent and requires higher level semantics, we can proceed by considering them as two mutually exclusive classes as defined by some oracle.

Considering the value observed at a pixel over time is usually resulted from different real world processes, a Gaussian mixture is appropriate to model the distribution, with each Gaussian representing an underlying process.

$$P(x) = \sum_{k=1}^{K} P(G_k) P(x \mid G_k) = \sum_{k=1}^{K} w_k \cdot g(x, \mu_k, \sigma_k)$$

G_k is the k-th Gaussian and $g_k(x) \equiv g(x, \mu_k, \sigma_k)$ is the normal density function. Under this distribution model, the original posterior probability can be reformulated as

$$P(B|x) = \sum_{k=1}^{K} P(B|G_k)P(G_k|x) = \sum_{k=1}^{K} P(B|G_k)[\frac{P(x|G_k)P(G_k)}{P(x)}]$$

$$= \frac{\sum_{k=1}^{K} P(x|G_k)P(G_k)P(B|G_k)}{\sum_{k=1}^{K} P(x|G_k)P(G_k)}$$

Segmentation consists of two independent problems: estimating the distribution of *all* observations at the pixel (within a time window) as a Gaussian mixture, and evaluating how likely each Gaussian in the mixture being background. There is an intuitive interpretation for this framework. Considering the observations at each pixel location are resulted from a number of discrete processes, we first color quantize them to reveal the underlying processes, then each process is evaluated as being foreground or background. Therefore, a pixel observed at a given time is classified as foreground if its probability of being foreground, expressed in terms of a mixture distribution and posteriors of the Gaussians, is greater than 0.5.

Solution to the first problem provides us with estimates of $P(G_k)$ and $P(x|G_k)$. Solution to the second problem gives us an estimate of $P(B|G_k)$. The first problem is analogous to color quantizing pixel values to reveal underlying processes, and is relatively application independent, at least from a theoretical standpoint. The second problem relates to classifying individual process as foreground or background, which is inevitably application dependent and heuristic based. This is where domain specific priors or higher level semantics can be imposed.

Obviously, perfect segmentation can not be achieved at the pixel level alone. For example, it would be impossible to distinguish a walking person from a rotating fan whose signal shares identical characteristics without an understanding of the event. There are a number of ways for imposing higher level semantics on background segmentation using region and frame level processing [9], active masks [4] or coupled object models [6]. It is difficult to make generalization on this type of interaction, and the topic is beyond the scope of this paper.

2.2 Background Model

Solving the two density estimation problems discussed in the previous section provides all the necessary information for performing background segmentation. Nonetheless, it is useful to define what *is* the background model. This problem has not been explicitly addressed in the literature, and it is often confused with the pixel distribution model. However, we show that there is a natural and theoretically sound definition under this framework.

If we can separate all observations into their respective classes, the background model should consist of the portion of observations that are believed to be background, or $P(x,B)$. Assuming $P(x|G_k,B)=P(x|G_k)$, the background model at pixel (r,c) at time t, $M(r,c,t)$, should be represented by

$$M(r,c,t) = P(x,B) = \sum_{k=1}^{K} P(x|B,G_k)P(G_k|B)P(B)$$

$$= \sum_{k=1}^{K} P(x|G_k)P(G_k)P(B|G_k)$$

In the most general form, the background model consists of the original mixture distribution $P(x)$ with the Gaussians weighted additionally by $P(B|G_k)$. However, the actual model depends on the assignment of $P(B|G_k)$. For instance, if we enforce only the best Gaussian candidate to have $P(B|G_k)=1$, and the rest 0, we still have a single Gaussian background model, although it is

selected from a mixture distribution. Similarly, this is a generalization of the labeling rule used in earlier works [4,8] where the best *b* Gaussians are labeled background. This is equivalent to a dichotomous decision with $P(B|G_k)$ equaling either 1 or 0, but the number of Gaussians labeled as background is not limited to one. However, a binary labeling leads to abrupt changes when a Gaussian switches from foreground to background or vice versa. Those discontinuities are eliminated in the generalized formula. We should point out that $M(r,c,t)$ incorporates $P(B)$, which expresses the probability of observing the actual background. This is useful in applications such as meeting systems where the background behind a person is never revealed.

It is often useful to obtain an image representation of the background model for the purpose of visualization or analysis of background events. A straightforward solution is to use the mean of the Gaussian most likely to be background. The drawback for this method is that the image will stay constant until a change suddenly occurs when a different Gaussian becomes the best candidate. A more intuitive representation is the expected value of the background process. Therefore,

$$E[x_{r,c}|B] = \sum_{k=1}^{K} E[x] \cdot P(x|G_k)P(G_k|B)$$

$$= \frac{\sum_{k=1}^{K} \mu_k(t) \cdot P(B|G_k)P(G_k)}{\sum_{k=1}^{K} P(B|G_k)P(G_k)}$$

The image is calculated as an average of the Gaussian means, weighted proportionally by their weights and posterior probabilities of being background. Another alternative is to use $E[x,B]$ and represent $P(B)$ through the alpha channel.

3. IMPLEMENTATION

The conceptual framework presented in Section 2 depends on $P(x)$ and $P(B|G_k)$. We describe our implementation for these two density estimation problems in this section.

3.1 Estimate P(x)

Using Gaussian mixture for density estimation is a well studied problem. Considering the real-time nature of video signals, the constraint for our application is that an *online*, instead of *batch*, learning algorithm is needed and the model must adapt to distribution changes over time. The adaptive filtering algorithm typically used [2,4,8] employs a fixed learning rate and converges very slowly. We propose using an adaptive learning rate schedule for each Gaussian that significantly improves the convergence speed and approximation results. We summarize the algorithm below. A detailed discussion and experimental results of this algorithm can be found in [5].

Let $w_i(t)$, $\mu_i(t)$, $\sigma_i^2(t)$ be the weight, mean and variance estimation of the i-th Gaussian at time t. The weights and means are initialized to 0. Variances are set to a large value V_0. A parameter α controls temporal retention. Then at time t, for any Gaussian G_k that matches x, its parameters are updated

$$p_k(x) = \frac{w_k \cdot g_k(x)}{\sum_{i=1}^{K} w_i \cdot g_i(x)}$$

$$c_k = c_k + p_k(x)$$

$$\eta_k = p_k(x) \cdot (\frac{1-\alpha}{c_k} + \alpha)$$

$$\mu_k(t+1) = (1-\eta_k) \cdot \mu_k(t) + \eta_k \cdot x$$
$$\sigma_k^2(t+1) = (1-\eta_k) \cdot \sigma_k^2(t) + \eta_k \cdot (x - \mu_k(t))^2$$

A point matches a Gaussian if it is less than 3 standard deviations away. If no Gaussian matches x, one of the Gaussian is reassigned.

$$k = \arg\min_i\{w_i\}$$
$$w_k = 0 \qquad \mu_k = x \qquad \sigma_k^2 = V_0 \qquad c_k = 1 \ .$$

After every iteration, all weights are updated using

$$w_i(t) = (1-\alpha) \cdot w_i(t-1) + \alpha \cdot (p_i(x) - w_i(t-1)) \ .$$

The criterion for selecting a Gaussian for reassignment can be application dependent. For example, we have also tried $\arg\min_i\{P(B|G_i)\}$ to keep good background candidates around longer, but we have noticed very little difference in performance. Compared to the work of [8] where $\eta_k = \alpha \cdot g_k(x)$, we observed the new formulation to perform much better.

3.2 Estimate P(B|G_k)

Unlike the previous problem of density estimation where the objective and desired algorithm behaviors are well defined, estimating $P(B|G_k)$ is largely based on heuristics and application dependent. However, compared to the original classification task of $P(B|x)$, the problem is simpler because more context can be utilized for estimating $P(B|G_k)$.

Since background is typically observed more often and displays less variation in value, w/α provides a good basis for the decision [8]. In addition, domain specific priors based on the location of the pixel or the mean of the Gaussian can be incorporated. For example, in our system for detecting people in meeting videos, a bias is placed against skin-colored Gaussians being background [6]. Furthermore, background models at neighboring locations and global statistics over the entire image provide additional context for refining this estimate.

In this implementation, we approximate $P(B|G_k)$ with $b_k = w_k \cdot (E_\sigma / \sigma_k)$ where E_σ is the expected value of the variance for a background Gaussian. We estimate this by averaging the variance of the top 25% of Gaussians in the entire image that have the largest $P(B|G_k)$. The maximum is set to 1.

It should be pointed out that $P(B|G_k)$ is considered independently for each Gaussian and, therefore, does not sum up to one. This discussion eventually depends on the precise definition of *background*. Nevertheless, this is the most general case and can be constrained to suit different applications. For instance, in most surveillance applications more interests are placed on moving objects. Any object can become background after dormant for a while. Under that definition, it is reasonable to assume a certain amount of observation during any time window belongs to background. This is the strategy used in [4,8] where the background process is estimated by selecting a minimum number of Gaussians to provide at least T percent coverage of the observations. Obviously, the same heuristic can be in our framework. On the other hand, in certain applications such as meeting video analysis, the true background for some location of the video can be constantly occluded and never revealed. This conclusion may come about from other evidence such as where the speaker is expected to be or a reliable object model which has identified the presence of a person in that location. In those situations, it may be desirable to indicate the fact that the "background" is never seen rather than calling the person background. Of course, how the true background can be estimated in that case is a different issue.

4. EXPERIMENTAL RESULTS

The proposed framework is tested on traffic and meeting videos. The same mixture density estimation and classification modules are used for both sequences. We used 3 Gaussian mixtures with $\alpha = 0.005$. Input is YUV space and diagonal covariance matrices were assumed. The RGB space produces similar but slightly more speckled results. Since the purpose of the experiment is to compare the proposed framework to prior arts, for which we used the algorithm described in [8], we did not include any domain specific priors such as skin tone detection for the meeting video or any special handling of shadows for the traffic sequence. The system runs at 15fps on 160x120 video on a 2GHz PC.

The effectiveness of the proposed method can be seen from Figure 1, which shows the results on four frames roughly 15 seconds apart from the traffic video. Results on the second row show that the prior art was unable to separate out the road from a relatively heavy traffic, and picked out only high contrast area. The reason is that the prior art adapts very slowly and requires the road to remain empty for a sufficiently long time to construct an accurate background model. As a result, the road and oncoming traffic were merged into a Gaussian with a large variance. In contrast, our proposed method was able to quickly estimate the true color and illumination variance of the road and detect the entire vehicle, as shown in the bottom row. Independently of the framework, these results can be further improved had shadow removal been done.

Similar observations can be made from the meeting video in Figure 2. In these three frames taken approximately 20 seconds apart, the prior art method was able to pick out the person only when he moved to a high contrast area. In the middle frame, the person was completely missing and blended to the background. On the other hand, the proposed method achieved almost perfect segmentation except in the last frame where part of his body was missing because the shirt color matched the dark background behind. This is a limitation that can not be overcome with this approach without high level information.

The algorithms' ability to adaptation is contrasted in the background models shown at the bottom two rows. The mean of the background Gaussian produced by the labeling rule [8] is shown in the fourth row, and by the expected value of the mixture as proposed in Section 2 is shown in the fifth. From the first column, which is 20 seconds into the video, it can be seen that the prior art largely maintained the initial value of the background. The person's original position at time zero was clearly visible. The proposed method, however, obtained a good estimate of the room without a person although the room has never been empty. Half way between the first column and the second column, the mug was removed. As the background model began to shift, it can be seen from the middle column that the old method made the transition by leaving out fragments of the mug; whereas in the proposed framework, the transition was made by fading out the mug. Finally, in the last column, the old method still had a pretty poor estimate of the background after one minute, but the proposed method had constructed a new background model without the mug.

5. CONCLUSIONS

We have presented a statistical framework for background segmentation based on Gaussian mixture modeling. We showed that a set of intuitive and theoretically sound solutions can be formulated in terms of two density estimation problems. With our proposed solution to those problems, the framework was applied to meeting and traffic video segmentation. The superior performance over existing methods validates our theory.

6. REFERENCES

[1] A. Elgammal, D. Harwood and L. Davis, "Non-parametric model for background substraction," *ECCV,* v.2, pp.751-767, 2000.

[2] N. Friedman, S. Russell, "Image segmentation in video sequences: a probabilistic approach," *Proc. 13th Conf. Uncertainty in Artificial Intelligence*, Aug. 1997.

[3] I. Haritaoglu, D. Harwood, and L.S. Davis, "W4: Real-time Surveillance of People and Their Activities," *PAMI,* 22(8), pp. 809-830, Aug. 2000.

[4] M. Harville, "A Framwork for High-Level Feedback to Adaptive, Per-Pixel, Mixture-of-Gaussian Background Models," *ECCV*, pp.543-560, May 2002.

[5] D.S. Lee, "Improved Adaptive Mixture Learning for Robust Video Background Modeling," *MVA*, pp. 443-446, Dec. 2002.

[6] D.S. Lee, J. Hull, B. Erol, "Segmenting People in Meeting Videos Using Mixture Background and Object Models," *Pacific-Rim Conf. on Multimedia*, pp. 791-798, Dec. 2002.

[7] N. Ohta, "A Statistical Approach to Background Subtraction for Surveillance Systems," *ICCV*, pp.481-486, July 2001.

[8] C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," *CVPR*, v.2, pp.246-252, June 1999.

[9] K. Toyama, J. Krumm, B. Brumitt, B. Meyers, "Wallflower: Principles and Practice of Background Maintenance," *ICCV*, pp. 255-261, Sept. 1999.

[10] C. Wren, A. Azarbayejani, T. Darrel and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *PAMI*, 19(7), pp.780-785, July 1997.
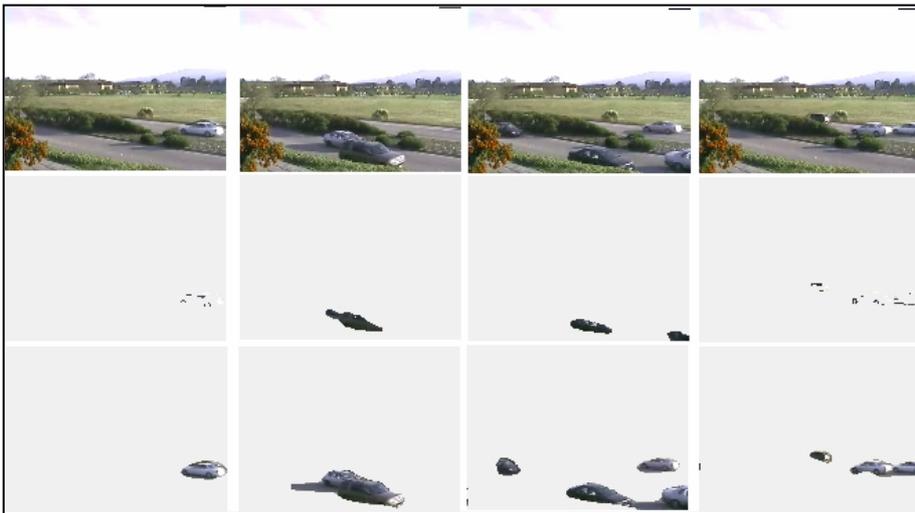
Figure 1 – From left to right, background subtraction results for frames 155, 485, 640 and 865 in the traffic sequence. The top row shows the original frames. The method described in [8], shown in the second row, adapted too slowly and picked out only high contrast regions. Results obtained using the proposed method, shown at the bottom, was able to detect moving cars very early on. No shadow removal was applied.
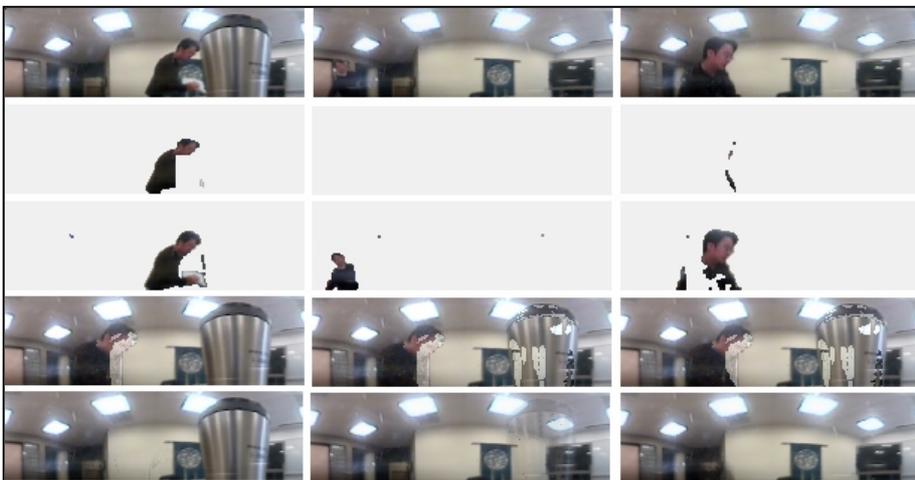
Figure 2 – Segmentation results for frames 660, 1230, 1810 from the meeting video are shown left to right. The top row shows the original frame. The second row shows results from a prior art method, which was unable to detect the person other than in high contrast area. The proposed method, shown in the third row, achieved much better segmentation. Background models produced by the labeling rule and the proposed formulation are shown in fourth and fifth row, respectively.