Department of Electrical
and
Computer Systems Engineering

Technical Report
MECSE-8-2004

# A Re-Evaluation of Mixture-of-Gaussian Background Modeling

H. Wang and D. Suter

MONASH
UNIVERSITY

# A RE-EVALUATION OF MIXTURE-OF-GAUSSIAN BACKGROUND MODELING

*Hanzi Wang and David Suter*

Institute for Vision Systems Engineering
Department of. Electrical. and Computer Systems Engineering
Monash University, Clayton 3800, Victoria, Australia
{hanzi.wang; d.suter}@eng.monash.edu.au

## ABSTRACT

Mixture of Gaussians (MOG) has been widely used for robustly modeling complicated backgrounds, especially those with small repetitive movements (such as leaves, bushes, rotating fan, ocean waves, rain). The performance of MOG can be greatly improved by tackling several practical issues. In this paper, we quantitatively evaluate (using the Wallflower benchmarks) the performance of the MOG. with and without our modifications. The experimental results show that the MOG, with our modifications, can achieve much better results - even outperforming other state-of-the-art methods.

## 1. INTRODUCTION

Background modeling is an important and fundamental part for many vision tasks such as real-time motion segmentation, tracking, video/traffic surveillance and human-machine interface.

In recent years, many background models have appeared [1-9]. Pfinder [6] is built upon the assumption that the scene is less dynamic than the object to be tracked and that the background is distributed according to a single Gaussian distribution. Although Pfinder can deal with small or gradual changes in the background, it fails when the background scene involves large or sudden changes, or has multi-modal distributions (such as small repetitive movements). The $W^4$ system [9] modeled the background scene by maximum and minimum intensity values, and the maximum intensity difference between consecutive frames in training stage. However, the background model from $W^4$ may be inaccurate when the background pixels are multi-modal distributed or widely dispersed in intensity.

The pixel-level Mixture of Gaussians (MOG) background model has become very popular because of its efficiency in modeling multi-modal distribution of backgrounds (such as waving trees, ocean waves, light reflection, etc), its ability to adapt to a change of the background (such as gradual light change, etc.) and the potential to implement the method in real time. Friedman and Russell [10] modeled the intensity values of a pixel by using a mixture of three Normal distributions and applied the proposed method to traffic surveillance applications. Stauffer and Grimson [4] presented a method that models the pixel intensity by a mixture of K Gaussian distributions.

Although many variants of the MOG background model [4, 5, 11] have been proposed, and MOG has been reported as being used in a wide variety of the systems (e.,g., for tracking [6, 7, 12], traffic surveillance [10], etc.), few papers provide a quantitive evaluation of the MOG method for background modeling. Toyama et. al. [1] implemented MOG and compared the result of MOG with that of "Wallflower", claiming superiority of the latter. In this paper, we show that the result of MOG can be greatly improved if we modify the implementation of MOG in some aspects: dealing with shadow removal, background update, background subtraction. This provides a re-evaluation of MOG using the same set of benchmarks as used in Wallflower study.

## 2. MIXTURE OF GAUSSIAN MODEL

In this section, we briefly describe the MOG model.

The basic idea is to assume that the time series of observations, at a given image pixel, is independent of the observations at other image pixels. It is also assumed that these observations of the pixel can be modeled by a mixture of K Gaussians (K is usually set from 3 to 5). Let $x_t$ be a pixel value at time $t$. Thus, the probability that the pixel value $x_t$ is observed at time $t$ is [4]:

$$P(x_t) = \sum_{i=1}^{K} \frac{w_{i,t}}{(2\pi)^{\frac{n}{2}} \left|\Sigma_{i,t}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_t-\mu_{i,t})^T \Sigma_{i,t}^{-1}(x_t-\mu_{i,t})} \quad (1)$$

where $w_{i,t}$ is the weight, $\mu_{i,t}$ is the mean value, and $\Sigma_{i,t}$ is the covariance matrix for the $i$th Gaussian distribution at time $t$.

For computational reasons, each channel of the color space is assumed to be independent from the other channels. The covariance matrix can then be written as:

$$\Sigma_{i,t} = \sigma_k^2 I \quad (2)$$

The K distributions are sorted by $w/\sigma$ and only the first B distributions are used to model the background, where

$$B = \arg\min_b \left( \frac{\sum_{i=1}^{b} w_i}{\sum_{i=1}^{K} w_i} > T \right) \quad (3)$$

$T$ is a threshold for the minimum fraction of the data used to model the background.

## 3. SOME PRACTICAL ISSUES

In a realistic environment, we find that using the MOG model is not enough to solve all problems met in background modeling. For example, a moving shadow region may be wrongly marked as foreground due to the illumination change, or relocation of a background object may result in some pixels in both the new and previous position of the background object being wrongly

labeled as foreground pixels, or a quick illumination change such as light switched on/off will greatly change the color of the background and increase the number of falsely detected foreground pixels, etc.

### 3.1 Shadow removal

Incorrectly labeling shadows as foreground pixels may cause failure in applications such as tracking, video surveillance, motion segmentation, etc.

When shadows appear or disappear, it is usually assumed that the chromaticity part at the pixel is not significantly changed. Normalized color is used in many background modeling methods such as [5, 7, 8] because normalized color is robust and less sensitive (than RGB color) to small changes in illumination caused by shadows.

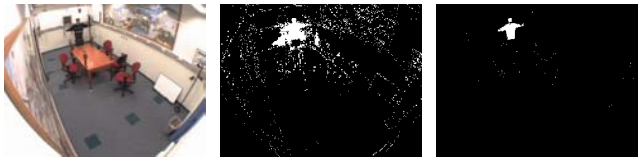
The normalized chromaticity coordinates can be written as:

$$\begin{aligned} r &= R/(R+G+B) \\ g &= G/(R+G+B) \\ b &= B/(R+G+B) \end{aligned} \tag{4}$$

Although using chromaticity coordinates can suppress shadows, the intensity information will be lost. Thus, we adopt the feature space ( $r$, $g$, $I$) as in [8], where $r$, $g$ are scaled to the range [0, 255] (assuming the 8 bit image values are used).

Let ($r_b$, $g_b$, $I_b$) be the expected value of a background pixel and ($r_t$, $g_t$, $I_t$) be the observed value at this pixel in frame $t$. If the background is totally static, we can expect $\beta \le I_t / I_b \le 1$ when the pixel is covered by shadow and $1 \le I_t / I_b \le \gamma$ when the pixel is highlighted by strong light. Figure 1 (c) shows an example where the shadow of the person is suppressed when using chromaticity coordinates ($r$, $g$) and the criterion that the intensity $I$ is such that $\beta \le I_t / I_b \le \gamma$ (compare with figure 1(b) RGB).

However, the background may be dynamic, i.e., multi-modal distributed. Let $\mu_i$ be the mean value and $\sigma_i$ be the standard variance of the $i$th Gaussian distribution. For the $i$th Gaussian distribution, we replace $I_b$ in the above criterion with the mean value $\mu_i$ (that is: $\beta \le I_t / \mu_i \le \gamma$ ).



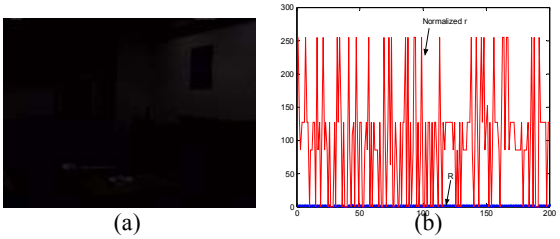
(a) (b) (c)

Fig. 1. (a) Image of a person and shadows; (b) Detection result using RGB; (c) using ($r$, $g$, $I$).

Another problem is that when the intensity is low, the normalized color ($r$ or $g$) is very noisy. Consider the image sequence "Time of Day" (TOD) in the Wallflower dataset, which displays a room gradually changing from dark to bright. In the first several hundred frames, the intensities of image pixels are very low. Figure 2 (b) shows the distribution of pixel values of the $r$ channel in the normalized color space and the R channel of RGB color space at image pixel (1, 1) in the first 200 frames. The stand variance of the pixels values in the $r$ channel, at image pixel (1, 1) for the first 200 frames, is 81.97; while the stand variance of the pixels values in the R channel is 0.91. To solve this problem we express the values of a pixel $x$, we use a mixed color space:

$$x = \begin{cases} (r,g,I) \ if \ I \ge Itd \\ (R,G,I) if \ I < Itd \end{cases} \tag{5}$$

where $Itd$ is a threshold. This modification improves the results, especially for video sequences including dark scenes, of background modeling (see section 4 for the results).




(a) (b)

Fig. 2. (a) The first frame of TOD; (b) The distributions of pixel values in the normalized $r$ channel and R channel of the RGB color space at image pixel (1,1) in the first 200 frames of TOD.

### 3.2 Updating the Background

Following [4], given a new observation $x_t$ that belongs to the $i$th Gaussian distribution, the parameters of the $i$th Gaussian distribution at time $t$ are updated as follows:

$$\mu_{i,t} = (1-\alpha)\mu_{i,t-1} + \alpha x_t \tag{6}$$

$$\sigma^2_{i,t} = (1-\alpha)\sigma^2_{i,t-1} + \alpha(x_t - \mu_{i,t})^T(x_t - \mu_{i,t}) \tag{7}$$

the weight of the $i$th Gaussian distribution is adjusted as follows:

$$w_{i,t} = (1-\alpha)w_{i,t-1} + \alpha \mathrm{M}_{i,t} \tag{8}$$

where $\alpha$ is learning rate; $\mathrm{M}_{i,t}$ is 1 when the new observation matches the $i$th distribution, and 0 otherwise.

This mechanism of updating the background has several advantages: such as robustly adopting to gradually light changing. However, if a background object is relocated to a new place, or if a new object is inserted into the background, the image pixels at both the new and previous position of the relocated background object or at the position of the inserted object, will not match the estimated K Gaussians and will be classified as foreground pixels. Although such changes of relocated or inserted background object may be temporarily of interest, it is not desirable to maintain these as foreground for a very long time. One common feature of the relocated or inserted background object is that once the position of the object is changed, typically, the object will stay there for a while. Thus, we employ a set of counters, which we call the "foreground support map"(FSM). FSM represents the number of times a pixel is classified as a foreground pixel:

$$FSM(x,t) = \begin{cases} FSM(x,t-1)+1 & if \ x \ is \ foreground \ pixel \\ 0 & if \ x \ is \ background \ pixel \end{cases} \tag{9}$$

When the FSM value of a pixel is larger than a threshold Ftd, we adopt this pixel to the background and use equations (6) – (8) to update the Gaussian model. This puts a time limit on how long a pixel can be considered as a static foreground pixel.

Another issue is that of choosing the learning rate $\alpha$. A high learning rate enables MOG to more quickly adapt to sudden

scene changes such as a light switching on/off, a sudden lightning, a sudden movement of uninteresting object, etc. However, a high learning rate also causes interesting foreground objects to quickly fade into the background. To obtain a satisfactory trade-off value is hard. Thus, we use extra information to adjust the learning rate. If the pixel number of detected foreground pixels is larger than a threshold (e.g., 70% of the whole image pixels as in Wallflower), we adjust the learning rate to a high value; otherwise, we set the learning rate to a low value.

### 3.3 Background Subtraction

Let $x_j$ be the $j$th component of pixel $x$. If $\left|x_{j=1or2} - \mu_i\right| > m\sigma_i$ is true for all $i$=1,…K ($m$ is usually set 2.5), or if $x_3 / \mu_i > \gamma \; or \; x_3 / \mu_i < \beta$ is true for all $i$=1,…K, the pixel is labeled as a foreground pixel.

However, there are two issues that should be considered: (a) the estimated standard variance could be overestimated or underestimated because the distribution of the pixels is not an ideal Gaussian. (b) when the intensity of a pixel is low, the value of $x_3 / \mu_i$ can be very varied even when the pixel belongs to the $i$th Gaussian. Thus, to solve issue (a), we set an upwards threshold $S_{max}$ and a downwards threshold $S_{min}$ for the estimated standard variance. $S_{max}$ and $S_{min}$ are respectively set to 0.1 and 15. To judge if the pixel is too far from the $i$th Gaussian, we check if $\left|x_j - \mu_i\right| > \max(m\sigma_i, \lambda)$ , where $\lambda$ is a threshold and is empirically set to 5. To solve issue (b), we use the criterion $x_3 / \mu_i > \gamma \; or \; x_3 / \mu_i < \beta$ for pixels with high intensities and $\left|x_3 - \mu_i\right| > \max(m\sigma_i, \lambda)$ for pixels with low intensities. Thus, we label a pixel as a foreground pixel if:

$$\begin{cases} for\ pixels\ with\ I \geq Itd : \left|x_{j=1or2} - \mu_i\right| > \max(m\sigma_i, \lambda), or \\ \qquad\qquad (x_3 / \mu_i > \gamma \; or \; x_3 / \mu_i < \beta)\ for\ all\ i = 1,...k \\ for\ pixels\ with\ I < Itd : \left|x_{j=1or2or3} - \mu_i\right| > \max(m\sigma_i, \lambda)\ for\ all\ i = 1,...k \end{cases} \quad (10)$$

## 4. EXPERIMENTAL RESULTS AND COMPARISONS

Toyama et. al. [1] benchmarked their algorithm "Wallflower" using a set of image sequences where each sequence presents a different type of difficulty that a practical task may meet. The performance is evaluated against hand-segmented ground truth. Two terms are used in evaluation: False Positive (FP) is the number of background pixels that are wrongly marked as foreground; False Negative (FN) is the number of foreground pixels that are wrongly marked as background.

A brief description of the Wallflower image sequences follows: **Moved Object (MO)** - A person enters into a room, makes a phone call, and leaves. The phone and the chair are left in a different position. **Time of Day (TOD)** - The light in a room gradually changes from dark to bright. Then, a person enters the room and sits down. **Light Switch (LS)** - A room scene begins with the lights on. Then a person enters the room and turns off the lights for a long period. Later, a person walks in the room, switches on the light, and moves the chair, while the door is closed. **Waving Trees (WT)** - A tree is swaying and a person walks in front of the tree. **Camouflage (C) -** A person walks in front of a monitor, which has rolling interference bars on the screen. The bars include similar color to the person's clothing. **Boostrapping (B) -** The image sequence shows a busy cafeteria and each frame contains people. **Foreground Aperture (FA) -** A person with uniformly colored shirt wakes up and begins to move slowly.

We have tested three different variants of MOG**. MOG 1** uses mixed color space (normalized rgb color space for pixels with high intensities and in RGB color space for pixels with low intensities). Thus, for a image pixel with high intensity, $x$ is expressed by ($r$, $g$, $I$); for a image pixel with low intensity, $x$ is expressed by ($R$, $G$, $I$); **MOG 2** uses normalized rgb color space; Each image pixel value $x$ is expressed by ($r$, $g$, $I$); **MOG 3** uses RGB color space. Each image pixel value $x$ is expressed by ($R$, $G$, $I$). In each we eliminated the foreground pixels whose 4-connected foreground pixels number less than 8.

From table 1, we can see that none of the methods achieve a lower value in both FN and FP for all seven image sequences. However, the modified MOG methods achieved best results in total error (TE) and total error excluding the light switch image sequence (TE*). In contrast to [1] we have shown that MOG, albeit with some modifications, *can* achieve high accuracy in background modeling. For the foreground aperture image sequence, Wallflower achieved the best result. However, the authors of [1] used a region-level processing as a post-processing step for Wallflower. In contrast, we did not use a region-level post-processing step. For the light switch image sequence, Wallflower used frames with both light on and light off in the training stage. In the training stage, we used only frames with light off.

## 5. CONCLUSION

The purpose of this paper is re-evaluate MOG in background modeling in the light of some simple modifications one can make to tackle real world problems. The modifications can make MOG competitive, if not superior, to many other methods - including Wallflower, in contrast to the conclusions reached by the proponents of that algorithm [1].

**Acknowledgements:** This work is supported by and ARC grant DP0452416. The work was carried out within the Monash University Institute for Vision Systems Engineering. This support is gratefully acknowledged.

## 6. REFERENCES

*1.K. Toyama, et al. "Wallflower: Principles and Practice of Background Maintenance," ICCV, Greece, p. 255-261 1999.*

*2.D. Kottow, M. Koppen and J. Ruiz-del-Solar. "A Background Maintenance Model in the Spatial-Range Domain," in 2nd Workshop on Statistical Methods in Video Processing, Prague, Czech Republic 2004.*

*3.M. Harville. "A Framework for High-Level Feedback to Adaptive, Per-Pixel, Mixture-of-Gaussian Background Models," ECCV, Copenhagen, Denmark: p. 543-560 2002.*

*4.C. Stauffer and W.E.L. Grimson. "Adaptive Background Mixture Models for Real-time Tracking," CVPR, p. 246-252 1999.*

*5.A. Elgammal, D. Harwood and L.S. Davis. "Non-parametric Model for Background Subtraction," ECCV, Dublin, Ireland: p. 751-767 2000.*

*6.C.R. Wren, et al., "Pfinder: real-time tracking of the human body," PAMI, 19(7): p. 780-785, 1997.*

*7.S.J. McKenna, et al., "Tracking Groups of People,"CVIU, **80**: p. 42-56, 2000.*

*8.A. Mittal and N. Paragios. "Motion-Based Background Subtraction using Adaptive Kernel Density Estimation," CVPR,, Washington, DC, p. 302-309, 2004.*

*9.I. Haritaoglu, D. Harwood and L.S. Davis, "W4: Real-Time Surveillance of People and Their Activities," PAMI, 22(8): p. 809-830, 2000.*

*10.N. Friedman and S. Russell. "Image Segmentation in Vedio Sequences: A Probabilitic Approach," in In Proc. Thirteenth Conf. on Uncertainty in Artificial Intelligence, San Francisco, CA: p. 175-181 1997.*

*11.O. Javed, K. Shafique and M. Shah. "A Hierarchical Approach to Robust Background Subtraction using Color and Gradient Information," in IEEE Workshop on Motion and Video Computing, Orlando 2002.*

*12.M. Xu and T.J. Ellis. "Illumination-invariant Motion Detection Using Colour Mixture Models," in Proc. British Machine Vision Conference: p. 163-172 2001.*

| Algorithm | ET | MO | TOD | LS | WT | C | B | FA | TE | TE* |
|---|---|---|---|---|---|---|---|---|---|---|
| MOG 1 | f. neg. | 0 | 597 | 1481 | 44 | 106 | 1176 | 1274 | 6581 | 4431 |
| | f.pos. | 0 | 358 | 669 | 288 | 413 | 134 | 41 | | |
| MOG 2 | f. neg. | 0 | 170 | 980 | 43 | 113 | 1174 | 998 | 7676 | 5644 |
| | f.pos. | 36 | 1671 | 1052 | 294 | 448 | 157 | 540 | | |
| MOG 3 | f. neg. | 0 | 839 | 1965 | 97 | 304 | 1498 | 2290 | 10538 | 7801 |
| | f. pos. | 0 | 29 | 772 | 388 | 1559 | 224 | 573 | | |
| Tracey LAB LP[1] | f. neg. | 0 | 772 | 1965 | 191 | 1998 | 1974 | 2403 | 12035 | 8046 |
| | f. pos. | 1 | 54 | 2024 | 136 | 69 | 92 | 356 | | |
| Mixture of Gaussian[2] | f. neg. | 0 | 1008 | 1633 | 1323 | 398 | 1874 | 2442 | 27053 | 11251 |
| | f. pos. | 0 | 20 | 14169 | 341 | 3098 | 217 | 530 | | |
| Bayesian decision[2] | f. neg. | 0 | 1018 | 2380 | 629 | 1538 | 2143 | 2511 | 31422 | 15603 |
| | f. pos. | 0 | 562 | 13439 | 334 | 2130 | 2764 | 1974 | | |
| Eigen-background[2] | f. neg. | 0 | 879 | 962 | 1027 | 350 | 304 | 2441 | 17677 | 16353 |
| | f. pos. | 1065 | 16 | 362 | 2057 | 1548 | 6129 | 537 | | |
| Wallflower[2] | f. neg. | 0 | 961 | 947 | 877 | 229 | 2025 | 320 | 11478 | 10156 |
| | f. pos. | 0 | 25 | 375 | 1999 | 2706 | 365 | 649 | | |

Table 1: Experimental results by different methods on Wallflower benchmarks. (Note 1 was reported in [2]; note 2 were reported in [1]).

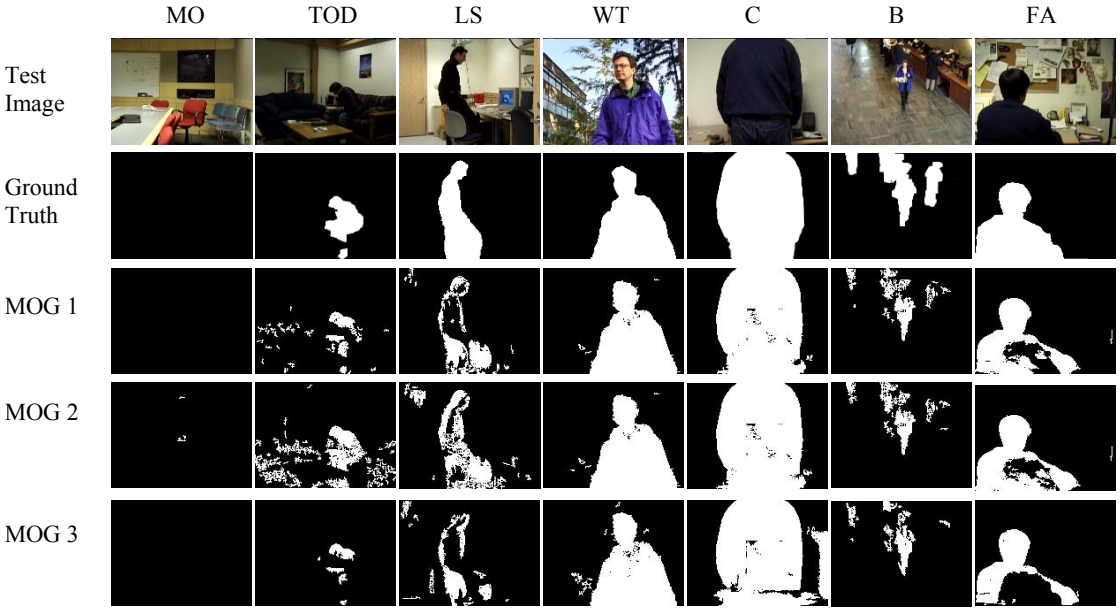



Fig. 3: Experimental results on the Wallflower benchmarks. The top row shows frames of each image sequences; the second row shows the hand-segmented ground truth; the third row to the fifth row show the results of three variants of MOG