Statistical Background Subtraction for a Mobile Observer

Eric Hayman and Jan-Olof Eklundh
Computational Vision and Active Perception Laboratory (CVAP)
Dept. of Numerical Analysis and Computer Science
KTH, SE-100 44 Stockholm, Sweden
{hayman, joe}@nada.kth.se
http://www.nada.kth.se/~hayman [*]

Abstract

Statistical background modelling and subtraction has proved to be a popular and effective class of algorithms for segmenting independently moving foreground objects out from a static background, without requiring any a priori information of the properties of foreground objects. This paper presents two contributions on this topic, aimed towards robotics where an active head is mounted on a mobile vehicle. In periods when the vehicle's wheels are not driven, camera translation is virtually zero, and background subtraction techniques are applicable. Parts of this work are also highly relevant to surveillance and video conferencing.

The first part of the paper presents an efficient probabilistic framework for when the camera pans and tilts. A unified approach is developed for handling various sources of error, including motion blur, sub-pixel camera motion, mixed pixels at object boundaries, and also uncertainty in background stabilisation caused by noise, unmodelled radial distortion and small translations of the camera.

The second contribution regards a Bayesian approach to specifically incorporate uncertainty concerning whether the background has yet been uncovered by moving foreground objects. This is an important requirement during initialisation of a system. We cannot assume that a background model is available in advance since that would involve storing models for each possible position, in every room, of the robot's operating environment. Instead the background model must be generated online, very possibly in the presence of moving objects.

1. Introduction

This paper discusses work aimed towards building a system for segmenting moving foreground objects from a static background for use with a mobile observer, such as a robot, equipped with a pan/tilt head. A hierarchy of algorithms can be employed, selected according to the camera's motion.

The simplest case is when the camera is static; neither the vehicle nor the active head moves. Background pixels maintain the same location in the image over time, and their depth in the scene is irrelevant. Next in complexity is the case when the camera rotates due to pan or tilt control signals to the active head. Provided the motion parameters can be obtained in some manner, they can be used to relate pixels in the current image to those of a reference image or mosaic. Thankfully, the depth of points in the scene is still irrelevant. A third, and much more complicated, situation arises when the camera also translates due to the vehicle being driven. The number of parameters increases vastly; depths in the scene must be computed, as must parameters for camera translation.

We present techniques for the first two cases listed above, motivated by the observation that a robot typically spends a significant proportion of its time with its wheels stationary. Our work is based on statistical background modelling and subtraction [4, 19, 15, 8, 21, 17]. The general idea is to obtain a per-pixel background model, and each pixel in a new image is examined to see whether it is feasible that it was drawn from that pixel's background model or not.

Although there is a great deal of literature on background modelling, and we shall shortly review some representative papers, there are some important issues which have received little attention previously, and which are crucial for satisfactory performance in our application. Hence,

- for pan/tilt heads we formulate a probabilistic, unified and efficient approach for coping with inaccuracies due to motion blur, mixed pixels at object boundaries, and errors in image stabilisation caused by noise, small camera translations or minor errors in calibration parameters such as focal length or radial distortion.

This algorithm is not just relevant to robotics, we also anticipate it being beneficial in other scenarios involving active cameras. We use a mechanism for mixed distributions previously derived by Kitamoto [11] for classifying aerial photography. To our knowledge this approach is novel within motion segmentation. We demonstrate a fast algorithm based largely on separable convolutions.

[*] Visit the website to see the movies accompanying this paper. This research was supported by the VISCOS project funded by the Swedish Foundation for Strategic Research, and also by a Marie Curie Fellowship of the European Community programme "Improving Human Research Potential" (Contract no. HPMFCT-2000-00650).

The second contribution concerns initialisation:

- we present a Bayesian classification algorithm which explicitly models the possibility that the background has not yet been uncovered by moving foreground objects.

The hypothesis that the background has not been seen is marginalised out to obtain the posterior probability that a given pixel in the current image belongs to the background. The evidence to support such hypotheses is gathered by comparing the compactness of distributions, and the length of time a stable component has been observed.

This is not an issue in, for instance, surveillance where one can accept either a separate, batch initialisation or online training with poor results until the model settles down. However, a mobile robot must be able to enter an unfamiliar environment and, once it has come to a stand-still, be able to provide a meaningful output as soon as possible, giving an honest estimate of how certain, or uncertain, the classification is. Thus initialisation cannot be considered a separate, offline process. Furthermore, it is not feasible to store background models for every location within the operating environment, instead they must be obtained online.

The remainder of this paper is organised as follows. Section 2 provides a description of previous work on background modelling and subtraction, placing particular emphasis on the algorithm of Stauffer and Grimson [19] on which we build. Section 3 introduces our technique for pan/tilt heads, while in Section 4 Bayesian reasoning is used to develop expressions for the posterior probabilities of background/foreground membership, incorporating the notion that the background might not yet have been observed. Section 5 presents experimental results whilst a critical discussion of our methods is provided in Section 6.

2. Previous work

In [12] Long and Yang suggest building a background model for independent motion segmentation as an alternative to image differencing between subsequent frames. More recently, there has been interest in algorithms with a sound statistical grounding [4, 19, 15, 8, 21, 17] with models which automatically adapt to the observed noise levels rather than requiring a threshold set manually. Most techniques are pixel-based: a model is built at each location in a reference image (or mosaic). The performance of background subtraction techniques can be characterised by the number of *false alarms* and *misdetections* [5].

There is a surprisingly small amount of literature which specifically deals with bootstrapping background models, and the assumption in almost all techniques is that the background was actually visible at some stage. One exception is a technique presented by Long and Yang in [12], but it is prohibitively costly. Another is a sequential algorithm due to Chien *et al.* [2], but there the background model is neither statistical nor adaptive. Gutchess *et al.* [6] use net optic flow in a neighbourhood around each pixel as an indicator that background is being uncovered. Yet even they require that the background is visible at some stage, and their technique is batch rather than sequential.

For moving cameras, Irani *et al.* [9] suggest a few potential blending schemes for creating static mosaics from dynamic scenes in the presence of camera motion. Median filtering is commonly used in the literature. In their work, no consideration is taken to noise levels varying either from pixel to pixel, or from sequence to sequence (without manual intervention at least). Other noteworthy publications within mosaicing for motion segmentation are [13, 10, 1].

Rowe and Blake [18] built a statistical background mosaic model for an active head. Mittal and Huttenlocher [14]) and Ren *et al.* [16] both extended Stauffer and Grimson's technique [19] to rotating cameras by incorporating a search for matching pixels within a region in the mosaic to accommodate registration errors. [16] weights the candidate locations according to a probabilistic spatial model.

2.1. Review of Stauffer and Grimson's algorithm

Our work uses the Gaussian Mixture Model (GMM) formulation developed independently by Stauffer and Grimson [19] and Friedman and Russell [4]. This scheme has an appealing statistical formulation, allows multimodal background models, and the background can change with time to accommodate slow lighting variations and objects blending into, or permanently leaving, the background.

The problem is formulated sequentially as follows: An image generation model $\mathbf{M}_t$ is assumed available for each pixel from previous measurements $\{\mathbf{Z}_0, \mathbf{Z}_1, ...\mathbf{Z}_{t-1}\}$. Given the current measurement $\mathbf{Z}_t$, the first aim is to determine whether this pixel was drawn from the background or foreground model. Subsequently, $\mathbf{Z}_t$ is used to update the model. Additionally, the model at $t = 0$ must be defined.

The image generation process is described by a GMM

$$\begin{aligned} P(\mathbf{Z}_t \mid \mathbf{M}_t) &= \sum_{h=1}^{H} \alpha_h \mathcal{N}(\boldsymbol{\mu}_h, \Sigma_h) \ , \\ \mathcal{N}(\boldsymbol{\mu}_h, \Sigma_h) &= \frac{1}{(2\pi)^{d/2} |\Sigma_h|^{1/2}} e^{-\frac{1}{2}(\mathbf{Z}_t - \boldsymbol{\mu}_h)^\top \Sigma_h^{-1} (\mathbf{Z}_t - \boldsymbol{\mu}_h)} \end{aligned}$$

where d is the dimensionality of the measurement space and each Gaussian is described by its mean $\boldsymbol{\mu}$ and covariance matrix Σ. The Gaussians are weighted by factors α_h where $\sum_h \alpha_h = 1$. $|\cdot|$ denotes the matrix determinant. The channels (RGB or YUV) are considered independent, that is

$$\Sigma_h = \begin{pmatrix} {}_1v_h & 0 & 0 \\ 0 & {}_2v_h & 0 \\ 0 & 0 & {}_3v_h \end{pmatrix} = \begin{pmatrix} {}_1\sigma_h^2 & 0 & 0 \\ 0 & {}_2\sigma_h^2 & 0 \\ 0 & 0 & {}_3\sigma_h^2 \end{pmatrix} \ .$$

In this paper left subscripts refer to the channel number, and will be dropped wherever possible.

This model contains one (or more) Gaussians to model the background, and the foreground is modelled by additional Gaussians. To gather the component(s) that correspond to the background first, the Gaussians are ordered by decreasing $(\alpha_h/_1\sigma_h)$, following the notion that these components not only explain much of the data (high α_h), but are also sharply peaked (low σ_h).

To determine the number of components C belonging to the background, a threshold T is specified such that

$$C = \arg\min_c \left(\sum_{h=1}^{c} \alpha_h > T \right) \tag{1}$$

This permits some pixels to have a multimodal background model to accommodate, for instance, monitor flicker, rotating fans or foliage of trees blowing in the wind.

A current measurement $\mathbf{Z}_t$, is assigned, if possible, to one of the components in the GMM by accepting the first component for which

$$\|Z_t - \mu_h\| < \kappa\sigma_h \tag{2}$$

for all channels. κ is a constant of the order 2–3. The label h' of this component determines whether the pixel is background or foreground in a binary segmentation map. The parameters of this component are then updated according to an exponential weighting scheme

$$\begin{aligned} \alpha'_h &\leftarrow (1-\delta)\alpha'_h + \delta \\ \mu'_h &\leftarrow (1-\delta)\mu'_h + \delta Z_t \\ \sigma'^2_h &\leftarrow (1-\delta)\sigma'^2_h + \delta(Z_t - \mu'_h)^\top(Z_t - \mu'_h) \end{aligned}$$

where δ determines the adaptation rate. For the other components $h \neq h'$ the mean and variance remain unchanged, but the α_h are modified according to $\alpha_h \leftarrow (1-\delta)\alpha_h$. [14] argues that constant weighting is preferable in some circumstances, such as when few observations are available.

If the test in eqn (2) fails for all components in the GMM, the pixel is labelled as foreground in the current image, and the least significant component of the GMM is replaced by

$$\alpha_h = \delta, \quad \mu_h = Z_t, \quad \sigma_h^2 = \bar{\sigma}^2$$

where $\bar{\sigma}^2$ is some initial high variance. This scheme is also used to initialise the model at time $t = 0$.

The scheme in eqn (2) for obtaining a *binary* segmentation map was chosen in [19] as an approximation to the true, Maximum A Priori (MAP) solution, to permit a real-time implementation (in 1999). For the MAP solution (indeed adopted in [16]) we introduce the discrete (boolean) variable A such that $A = A_1$ denotes that the pixel was drawn from the background distribution, and $A = A_2$ implies that the pixel belongs to the foreground. We seek the posterior probability $P(A_1 \mid \mathbf{Z})$ via Bayes' Rule

$$P(A_1 \mid \mathbf{Z}) = \frac{P(\mathbf{Z}|A_1)P(A_1)}{P(\mathbf{Z})} = \frac{P(\mathbf{Z}|A_1)P(A_1)}{\sum_j P(\mathbf{Z}|A_j)P(A_j)}$$

The likelihoods $P(\mathbf{Z}|A_1)$ and $P(\mathbf{Z}|A_2)$ are probability density functions (PDF's), integrating to unity. Assuming, for the sake of illustration, that eqn (1) dictates that $C = 1$,

$$\begin{aligned} P(\mathbf{Z}|A_1) &= \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1), \quad \text{and} \\ P(\mathbf{Z}|A_2) &= \frac{1}{\sum_{h=2}^{H} \alpha_h} \sum_{h=2}^{H} \alpha_h \mathcal{N}(\boldsymbol{\mu}_h, \Sigma_h) \end{aligned} \tag{3}$$

The remaining terms are clearly

$$P(A_1) = \alpha_1 \qquad \text{and} \qquad P(A_2) = \sum_{h=2}^{H} \alpha_h \ , \tag{4}$$

giving intuitive results for the products $P(\mathbf{Z}|A_j)P(A_j)$. The component $h'$of the GMM to update is that which maximises $\alpha_h \mathcal{N}(\boldsymbol{\mu}_h, \Sigma_h)$.

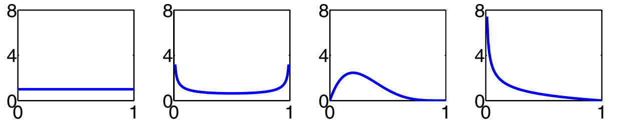
If all likelihoods $\mathcal{N}(\boldsymbol{\mu}_h, \Sigma_h)$ are very small the output will be somewhat arbitrary. This could be remedied by supplementing the GMM with a uniform distribution to account for a new foreground object appearing in view. Alternatively the original test of eqn (2) can be applied.

3. Background subtraction with an active head

Any scheme for background subtraction with a rotating camera must first compute the apparent background motion between the current image and the mosaic (reference image). This motion is characterised by an invertible one-to-one mapping $\mathcal{T}$ between pixel coordinates in the mosaic (x, y) and the current image (x', y'). Many papers have been written on the robust computation of this transformation, using either corner features or the intensity values. We assume that the calibration parameters (including radial distortion) are known reasonably accurately, leaving just two parameters, the pan and tilt angles. These are computed robustly from corner features with the MLESAC algorithm [20] using the mosaic to derive the reference image.

In principle, once this transformation has been found, the algorithm for statistical background modelling and subtraction can proceed largely as before. A larger lattice (mosaic) of GMM's than the original image is required since the aggregate field of view is wider. Pixels (x', y') in the new image can be classified using the GMM at location (x, y) in the mosaic. The GMM at location (x, y) is subsequently updated using the observation from (x', y').

However, we cannot expect motion estimation to be perfect due to image noise and imperfections in the geometric model caused by inaccuracies in the calibration parameters. With regard to foreground/background segmentation, geometric errors of just a single pixel could easily cause a pixel in the current image to be misclassified, merely because it is not being compared with the appropriate model, which is contained in one of the neighbouring locations in the lattice. The problem is exacerbated by the fact that the observed signal could be generated by a *mixture* of pixels in



$(\phi_1 = \phi_2 = 1)$ $(\phi_1 = \phi_2 = 0.5)$ $(\phi_1 = 2,\ \phi_2 = 5)$ $(\phi_1 = .5,\ \phi_2 = 2)$

Figure 1. Examples of beta distributions

the background model due to sub-pixel motions or motion blur. All these errors could be acute in areas of high texture.

3.1. Modelling mixed pixels

We now present a model which can successfully classify pixels in the presence of *all* these sources of error within a single mechanism. In all cases the observation $\mathbf{Z}$ at a pixel in the current image is a random process Z which was generated by a mixture of K random processes,

$$Z = \sum_{k \in R} a_k Z_k + E \tag{5}$$

where R is a region of K pixels around the location (x, y) in the mosaic of GMM's, and E is a noise term, independent of k and whose distribution is modelled as a zero mean Gaussian $\mathcal{N}(0, \Sigma_{\rm E})$. The variables a_k describe the area proportions of pixel k such that $0 \leq a_k \leq 1$ and $\sum_{k \in R} a_k = 1$. Z_k is a Gaussian from the GMM (more on this later).

The observation likelihood conditioned on the vector $\mathbf{a} = (a_1, a_2, ...a_K)^\top$ is given by the convolution of the individual PDF's

$$P(\mathbf{Z}|\mathbf{a}) = P(\mathbf{Z}|a_1) * ... * P(\mathbf{Z}|a_K) * P_{\rm E}(\mathbf{Z}) \tag{6}$$

where $*$ denotes convolution and $P_{\rm E}(\mathbf{Z})$ is the independent noise term. $\mathbf{a}$ can be eliminated by marginalization to yield the PDF of the so-called Mixel Distribution (MD)

$$P(\mathbf{Z}) = \int P(\mathbf{Z} \mid \mathbf{a})\ P(\mathbf{a})\ d\mathbf{a} \tag{7}$$

where $P(\mathbf{a})$ is the prior distribution of $\mathbf{a}$. We assume that the priors $P(\mathbf{a})$ are described by a $K - 1$ dimensional beta distribution (a Dirichlet distribution),

$$P(\mathbf{a} \mid \boldsymbol{\phi}) = \frac{\Gamma(\varphi)}{\prod_{k=1}^K \Gamma(\phi_k)} \prod_{k=1}^K a_k^{\phi_k - 1}$$

where Γ is the gamma function, the parameters $\boldsymbol{\phi} = \{\phi_1, \phi_2 ... \phi_K\}$ describe the shape of the distribution, and $\varphi = \sum_{k=1}^K \phi_k$. Examples are given in fig 1 for $K = 2$.

This type of model was previously used by Kitamoto [11] for classifying pixels from aerial images into classes such as cloud and sea. It has also been used with $K = 2$ in alpha matting [22] where the aim is to recover values of α_i at each pixel for superimposing fine-structured foreground objects (such as hair) onto a background.

The PDF in eqn (7) is computationally intractable and possibly multimodal. We settle for the same expedient as Kitamoto by approximating the PDF by its first two moments. Indeed, Kitamoto derived closed-form expressions for the mean and covariance of the mixel distribution as

$$\boldsymbol{\mu}_{\rm MD} = \frac{\sum_{k=1}^K \phi_k \boldsymbol{\mu}_k}{\varphi} \quad , \tag{8}$$

$$\Sigma_{\rm MD} = \Sigma_{\rm E} + \frac{\sum_{k=1}^K \phi_k(\phi_k + 1)\Sigma_k}{\varphi(\varphi + 1)} \ +$$

$$\frac{\sum_{k=1}^K \phi_k(\varphi - \phi_k)\boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top - \sum_{k=1}^K \sum_{f=1, f \neq k}^K \phi_k \phi_f \boldsymbol{\mu}_k \boldsymbol{\mu}_f^\top}{\varphi^2(\varphi + 1)}$$

We must clarify that the subscripts k denote the position in the region R, they do *not* refer to components within a pixel's GMM. $\Sigma_{\rm E}$ is the pixel-independent noise, as before.

The mean value of the MD takes an intuitive form as a weighted mean of the means of the individual random variables. Assuming that the covariance matrix is diagonal, some simple manipulation (which we omit here due to space limitations) of Kitamoto's expression leads to an expression for the variance ${}_l\sigma^2_{\rm MD}$ of each channel l

$${}_l\sigma^2_{\rm MD} = \frac{\sum_{k=1}^K \phi_k(\phi_k + 1)\ {}_l\sigma^2_k}{\varphi(\varphi + 1)} \ +$$

$$\frac{\sum_{k=1}^K \phi_k\ \ {}_l\mu^2_k}{\varphi(\varphi + 1)} - \frac{{}_l\mu^2_{\rm MD}}{\varphi + 1} + {}_l\sigma^2_{\rm E} \ . \tag{9}$$

From eqns (8) and (9) it is evident that the terms involving u_k can be computed by separable convolutions over the entire mosaic (or a region of interest therein): we assume that R is of fixed size $w \times w = K$ centred on the current pixel, and that $\{\phi_k\}$ can be described by a separable mask of size w. In our work we set $w = 7$ when using small (176×144) input images. $\{\phi_k\}$ is defined by the separable mask $\mathbf{b}\mathbf{b}^\top$ where $\mathbf{b} = (0.1, 0.13, 0.16, 0.18, 0.16, 0.13, 0.1)^\top$.

The mask $\{\phi_k(\phi_k + 1)\}$ in eqn (9) does not lead to a separable filter kernel, the matrix has rank 2, but for our choice of b (and all others we considered) the second singular value d_2 is at least a couple of orders of magnitude smaller than the first, d_1. The matrix is therefore well approximated by the closest matrix in Frobenius norm, which is given by setting $d_2 = 0$. A separable kernel is then obtained as $\sqrt{d_1}\mathbf{u}_1$ where $\mathbf{u}_1$ is the first left singular vector of the mask matrix.

In summary, both $\boldsymbol{\mu}_{\rm MD}$ and $\Sigma_{\rm MD}$, can be computed by sums of terms formed by separable convolution. The complexity of separable convolutions is *linear* in w, while the techniques in [16] and [14] scale as w^2.

3.2. Implementation details

Although this forms the basis of our approach, there are a few remaining issues to discuss. First, this argument has primarily been based on dealing with errors in *background* registration. We will therefore assume that all the first components, $h = 1$ in the region R of the GMM [1] represent the same layer (background or foreground), so it is the means and variances of this component which are input to eqns (8) and (9). Pixel classification proceeds as in Section 2.1, but now the first component of the GMM is replaced by the MD, $\boldsymbol{\mu}_1 \leftarrow \boldsymbol{\mu}_{\mathrm{MD}}$ and $\Sigma_1 \leftarrow \Sigma_{\mathrm{MD}}$.

If the random variables which comprise the MD were not drawn from the same layer, but from a mixture of background and foreground, the resulting observation model will tend to have a higher variance (assuming that inter-layer variation is larger than intra-layer variation). Thus while our scheme will reduce the false alarm rate, there is also an increased risk of misdetections. We are currently investigating the explicit incorporation of this notion in our Bayesian framework

The second matter concerns the class (pixel) independent noise term $\mathbf{E}$ in eqn (5). Until now we have modelled the variation within each class rather than the independent noise process, not that we have really had to distinguish between the two. Consider though what happens in a region R of uniform colour ($\boldsymbol{\mu}$ is constant) with identical class variances (Σ =const = $\bar{\Sigma}$) over the region R. If we neglect Σ_{E}, then Σ_{MD} is less than $\bar{\Sigma}$ unless $\boldsymbol{\phi}$, and thus also $\mathbf{a}$, has the form $\{0, 0, ...1, ...0\}$. Since this effect can be considerable, the model is clearly lacking in some respect. Thus, for each region and channel we identify some common noise variance ${}_l\sigma^2_{\mathrm{E}}$ and subtract it from each ${}_l\sigma^2_k$ before applying eqn (9). We select ${}_l\sigma^2_{\mathrm{E}}$ as the minimum ${}_l\sigma^2_k$ in the region R, ${}_l\sigma^2_{\mathrm{E}} = \min({}_l\sigma^2_1, ...{}_l\sigma^2_K)$, denoted ${}_l\sigma^2_{\min}$. This procedure does not incur a significant penalty on computation time. Eqn (9) can be rewritten as

$$\begin{aligned} {}_l\sigma^2_{\mathrm{MD}} &= \frac{\sum_{k=1}^{K} \phi_k(\phi_k+1)\,{}_l\sigma^2_k}{\varphi(\varphi+1)} + \beta\,{}_l\sigma^2_{\min} \\ &\quad + \frac{\sum_{k=1}^{K} \phi_k\ \ {}_l\mu^2_k}{\varphi(\varphi+1)} - \frac{{}_l\mu^2_{\mathrm{MD}}}{\varphi+1}\ , \qquad (10) \\ \beta &= 1 - \frac{\sum_{k=1}^{K} \phi_k(\phi_k+1)}{\varphi(\varphi+1)}\ , \qquad (11) \end{aligned}$$

which maintains the separable nature of the computations, and finding the minimum within a rectangular block is a common, cheap separable operation in morphological filtering. Nor is an additional pass through the data necessary.

A third issue concerns the further components of the GMM. Naturally, we could compute an MD in a similar manner by assuming that all components $h = 2$ of the GMM are drawn from the same object. Since foreground objects tend to be smaller than background objects, there is a very distinct risk that this assumption will be violated. Instead we assume that $\Sigma_{\mathrm{MD}} - \Sigma$ from the *first* component of the central pixel in R gives a fair estimate of the required increase in variance of the underlying random process which generated also the *further* component of the GMM. The mean of the original component is assumed unchanged.

Once a pixel has been classified as background or foreground, the GMM at location (x,y) is updated according to the scheme outlined previously in Section 2.1. This is different to [16, 14] who update the pixel within the region R which the current measurement best conforms with. Thus our mosaics could appear slightly more blurred than theirs.

Experimental results are presented in Section 5.

4. Dealing with covered background

Stauffer and Grimson's algorithm [19] described in Section 2.1 makes the fundamental assumption that the background has been seen, $C \geq 1$. If during the first few frames slowly moving foreground objects are present, these are immediately accepted as background, resulting in many misdetections. This is not intended to be a criticism of Stauffer and Grimson's algorithm. In their application with a surveillance camera which operates over many days, they are quite prepared to accept that the output becomes reliable only after a certain period of time. It is our requirement of a meaningful output as soon as possible, which motivates a new algorithm. Furthermore, we anticipate that a mobile robot will observe not just people who are walking, but also people standing or sitting relatively still, possibly gesturing to the robot to give it commands. We cannot expect our algorithm to segment out foreground pixels in uniformly coloured regions if the background is never revealed at that pixel's location. However, there is hope if the object is textured and moves slightly from side to side. Intuitively, pixel-based information is still available via frame-differencing.

To achieve these goals, we permit only a unimodal background model during initialisation. In the future we intend to investigate automatically relaxing this requirement once we are reasonably sure that the background has been seen.

4.1. Pixel classification

We now describe our novel approach to incorporate the notions (i) that the background might not yet have been seen, and (ii), if it has been seen we are not sure which component of the GMM contains it. Thus we supplement the model of Section 2.1 with a discrete random variable B with possible values $\{B_1, B_2, ..., B_{H+1}\}$ as follows

$B = B_1$	The background has been observed and is correctly identified as such in the most significant part of the GMM.
$B = B_i$ $i \in \{2, ...H\}$	The background *has* been seen, but is contained in the ith component of the GMM. The first component represents foreground.
$B = B_{H+1}$	The background has not yet been visible.

[1] assumed sorted by decreasing α/σ as described in Section 2.1

H is the number of Gaussians in the GMM, as previously.

The posterior probability of membership, conditioned on $\mathbf{Z}$ and also on some model parameters $\mathbf{M}$ is given by applying the sum and product rules of probabilities as well as Bayes' rule:

$$\begin{aligned} P(A_1|\mathbf{Z},\mathbf{M}) &= \sum_i P(A_1, B_i|\mathbf{Z},\mathbf{M}) \\ &= \sum_i P(A_1|B_i,\mathbf{Z},\mathbf{M})P(B_i|\mathbf{Z},\mathbf{M}) \\ &= \sum_i \frac{P(\mathbf{Z}|A_1,B_i,\mathbf{M})P(A_1|B_i,\mathbf{M})P(B_i|\mathbf{Z},\mathbf{M})}{\sum_j P(\mathbf{Z}|A_j,B_i,\mathbf{M})P(A_j|B_i,\mathbf{M})} . \end{aligned} \quad (12)$$

B_i is clearly independent of the current measurement, so $P(B_i|\mathbf{Z},\mathbf{M}) = P(B_i|\mathbf{M})$. In our current implementation the evidence $\mathbf{M}$ consists of two components. Following on from the philosophy introduced when ordering the components in the GMM in Section 2.1, the recovered variance is clearly of some use, as is the number of times N_i each component has been seen, thus $\mathbf{M} = \{\mathbf{\Sigma}, \mathbf{N}\}$ where $\mathbf{\Sigma} = \{\Sigma_1, ..., \Sigma_H\}$ and $\mathbf{N} = \{N_1, ..., N_H\}$.

Applying Bayes' rule yields

$$P(B_i \mid \mathbf{\Sigma}, \mathbf{N}) = \frac{P(\mathbf{\Sigma} \mid B_i, \mathbf{N})\, P(B_i \mid \mathbf{N})}{\sum_i P(\mathbf{\Sigma} \mid B_i, \mathbf{N})\, P(B_i \mid \mathbf{N})}$$

where $\sum_i P(B_i \mid \mathbf{N}) = 1$. Assuming that the covariance matrices Σ_i in the GMM are independent for different i,

$$P(\mathbf{\Sigma} \mid B_i, \mathbf{N}) = \prod_{i'}^{H} P(\Sigma_{i'} \mid B_i, \mathbf{N}) \ .$$

The key point is that if $i' = i$ this component corresponds to background, and we expect a low signal variance. On the other hand, for $i' \neq i$ we expect a higher variance corresponding to a foreground object. Since we have very little past history, we lack a *pixel-based* reference for what constitutes a viable background distribution, and we must instead settle for adopting a more *global* measure. For each channel, l, the median variance ${}_l v_{\text{med}}$ is computed over the entire image from the first components $h = 1$ of the GMM's. For this variable to be robust at least 50 % of the pixels must either be background or covered by a moving, but uniformly coloured, foreground object, in which case a purely pixel-based method is doomed to fail regardless. For each $i \in \{1, 2, ...H\}$ we model the likelihood of the variance Σ_i conditioned on B_i as an exponential PDF,

$$P(\Sigma_{i'}|B_i,\mathbf{N}) = \frac{1}{{}_1\lambda_{i'}\ {}_2\lambda_{i'}\ {}_3\lambda_{i'}} e^{-(\frac{{}_1 v}{{}_1\lambda_{i'}} + \frac{{}_2 v}{{}_2\lambda_{i'}} + \frac{{}_3 v}{{}_3\lambda_{i'}})}$$

In general the parameters ${}_1\lambda_i, {}_2\lambda_i, {}_3\lambda_i$ are not equal since the channels have different noise characteristics. For clarity of presentation we concentrate on a single channel and drop the left subscript. We define one set of parameters, λ_i, for $i' = i$ and another set, $\lambda_{i'}$, for $i' \neq i$ by requiring that the PDF's for each channel intersect at some multiple γ of the global, median variance

$$\frac{1}{\lambda_{i'}} e^{-\frac{\gamma v_{\text{med}}}{\lambda_{i'}}} = \frac{1}{\lambda_i} e^{-\frac{\gamma v_{\text{med}}}{\lambda_i}} \ .$$

This gives one constraint on $\lambda_{i'}$ and λ_i. A second is provided by specifying the ratio, r^* of the PDF's at $v = 0$,

$$\frac{1}{\lambda_i} = r^* \frac{1}{\lambda_{i'}} \ .$$

For reliance on $\mathbf{N}$, we adjust the r^* parameter as

$$r_i \leftarrow (r^* - 1)\rho_i + 1, \quad \rho_i = \min(1, N_i/N_{\max})$$

where $N_{\max}$ is a time constant. The parameters r^* and γ are also provided by the user. We use $r^* = 4$ and $\gamma = 4$.

For $i \in \{1, 2, ...H\}$ we let $P(B_i \mid \mathbf{N}) = N_i/(2N_{\text{tot}})$, and setting $P(B_{H+1} \mid \mathbf{N}) = 1/2$ then gives equal *total* prior probability to the hypotheses that the background has been seen $\{B_1...B_H\}$ or not $\{B_{H+1}\}$.

There are still some terms remaining in eqn (12). The terms $P(\mathbf{Z}|A_1, B_i, \mathbf{M})$ follow the expressions given in Section 2.1, and contain a single component i from the GMM. Similarly, the terms $P(\mathbf{Z}|A_2, B_i, \mathbf{M})$ correspond to the remaining terms in the GMM. Since we have limited faith in the recovered model during early frames, we additionally apply a Gaussian hyper-prior with variance Σ^* to the mean of each component of the GMM, which by the convolution properties of Gaussians simply leads to the variance of the original Gaussian being replaced as $\Sigma_i \leftarrow \Sigma_i + \Sigma^*$ where we choose Σ^* as $\Sigma^* = (1/\rho_i - 1)\Sigma_i$. Additionally we introduce a uniform component $p = 1 - \rho_i$ which during the first frames accounts for most of the data, but subsequently vanishes. Assuming the signal lies in the range [0,1]

$$P(\mathbf{Z}|A_1, B_h, \mathbf{M}) = (1 - p)\mathcal{N}(\boldsymbol{\mu}_h, \Sigma_h) + p \ .$$

For $i \in \{1, 2, ...H\}$ we set

$$P(A_1|B_i,\mathbf{M}) = \frac{1}{2} + \left(\frac{N_i}{N_{\text{tot}}} - \frac{1}{2}\right)\rho_i, \qquad (13)$$

where $N_{\text{tot}} = \sum_{i=1}^{H} N_i$. This implies that we use constant rather than exponential weighting in these early frames, following [14]. We define $P(A_1|B_{H+1}, \mathbf{M}) = (1 - \rho_{H+1})/2$ where $\rho_{H+1} = \max(1, N_{\text{tot}}/N_{\max})$. By the sum rule, $P(A_2|B_i, \mathbf{M}) = 1 - P(A_1|B_i, \mathbf{M})$.

4.2. Implementation details

The complete algorithm largely follows that outlined in Section 2.1. Our contribution lies in replacing the pixel classification stage. As stated above, we update the GMM's

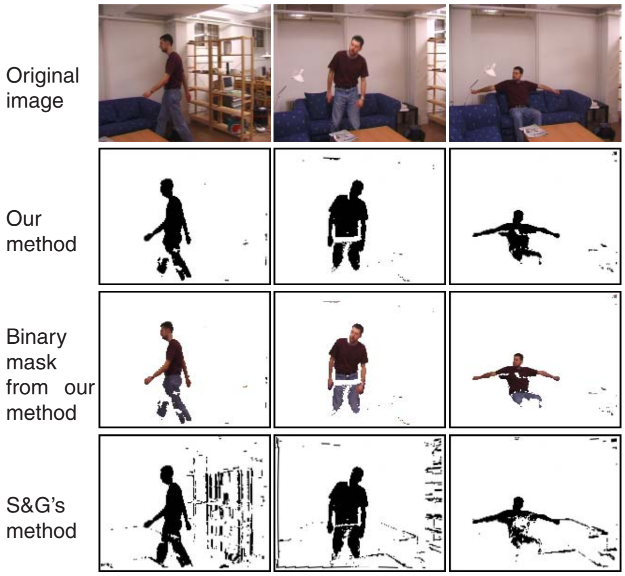


Figure 2. Segmentation results from a sequence taken with a rotating camera, comparing our algorithm with that in [19].

using constant rather than exponential weighting. We select which component of the GMM to update using the MAP scheme mentioned in Section 2.1. The algorithm requires three additional parameters, r^*, γ and the time constant $N_{\max}$, to be set manually.

With a moving camera, the most likely reason for observing high variances in the GMM is frequently *not* that the background has not yet been seen, but sub-pixel motion in regions of high texture. Therefore, the approach presented here is, as yet, unsuitable with rotating cameras.

5. Experiments

The algorithms presented in this paper were tested on various sequences using different cameras. Both methods run at 10 Hz on 176 × 144 images on a 2GHz P4 laptop.

Experiments with rotating cameras. The algorithm of Section 3 for pan/tilt heads was tested on a sequence obtained with a DV camcorder. Fig 2 demonstrates the results and compares them with the algorithm of Stauffer and Grimson[19]. The sequence is interesting in that there are both areas of highly and little cluttered background. Our algorithm performs well. Relative to Stauffer and Grimson's technique there are few false alarms, but misdetections can arise, as anticipated. For instance, misdetections can occur when skin occludes the wooden bookshelf in textured areas, due to their similar colour. Note that results are good where the blue jeans of the subject cover the blue sofa.

Experiments with static cameras. The algorithm presented in Section 4 for reasoning about whether the background has been seen or not, is tested on the "silent" MPEG-4 test sequence. The results in fig 3 are very satisfactory. Initially the MAP output from our method indicates a great amount of uncertainty, but as more data is gathered, the model matures and the output becomes more certain. Our algorithm successfully segments out the entire head, and indicates that the torso is uncertain throughout the entire sequence. Fig 3 also compares our technique with the method of [19] which unsurprisingly accepts more of the person as background. On the other hand, our algorithm is more sensitive to shadows: more of the contour of the lady's shadow is segmented out as foreground in the right of the image.

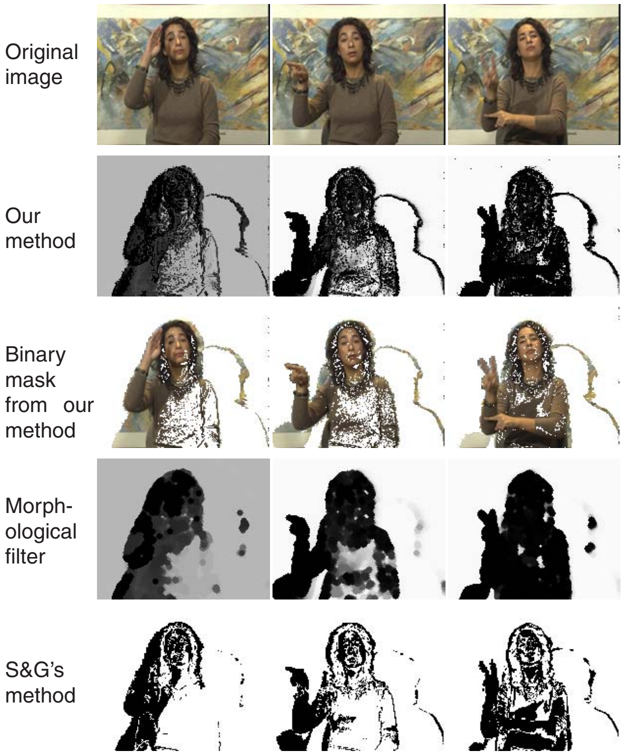


Figure 3. Segmentation results from three frames of the "silent" sequence. The top row shows the original sequence, the second the MAP output of the algorithm of Section 4. Dark areas indicate foreground, white background, and grey areas cannot be labelled with any degree of certainty. A binary segmentation mask derived from the MAP output is shown in the third row, while applying greyscale morphological filters to the MAP output yields the fourth. For comparison the binary output from Stauffer and Grimson's method is also shown.

We also demonstrate improvements possible using morphological filters in a post-processing stage. However, such filtering inevitably removes some of the finer detail; how much depends on the size of the structure element.

We briefly present results for further sequences in fig 4. We conclude that our technique works well for images with high signal to noise ratio ("Mother and daughter" sequence), and also performs slightly better than [19] with poor quality inputs ("Lab" sequence) when noise tends to

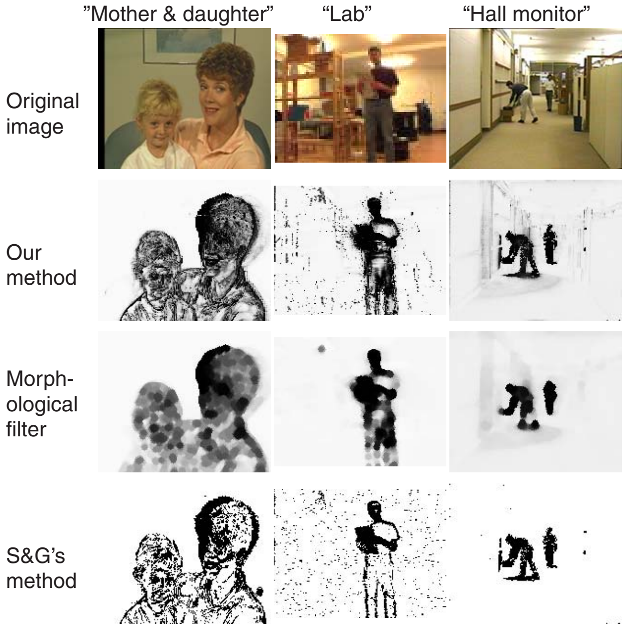


Figure 4. Applying the algorithm of Section 4 to a variety of sequences. The results are discussed in the main text.

be large in comparison with signal variations from slight motions of a rather uniformly coloured foreground. With the “hall monitor” sequence, the input data is also rather noisy. This example is merely included to illustrate that the extra machinery introduced by our algorithm does not have a destabilizing effect in cases where the background is indeed visible at some stage in the sequence.

6. Discussion and conclusions

This paper presented two novel algorithms for classifying pixels in new images into foreground or background by detecting independent motion relative to a statistical model of the background appearance. The techniques were primarily motivated by applications in mobile robotics.

The first algorithm was developed for panning and/or tilting active heads and gave a unified approach to handling motion blur, inaccurate motion estimation, geometric calibration errors, and mixed pixels at motion boundaries. Despite approximating an intractable PDF by its first two moments, satisfactory results were obtained. Separable convolutions permit an efficient implementation, making it faster than previous approaches. In addition to use in mobile robotics, we anticipate the method will be of great value also in video surveillance and conferencing.

The second technique specifically modelled the notion that the background might not yet have been uncovered during early stages of operation. To this end we examined the variances of the GMM’s components. This limited the algorithm to use with high signal-to-noise cameras on a static head. In the future this requirement could be relaxed by incorporating, into the same framework, further information obtained from a *neighbourhood* around the pixel. Net flow as in [6] or block matching as in [15] are natural candidates.

There is still much work required before our system can be termed complete. Robust operation requires mechanisms for handling shadows and rapid, global changes in illumination. This can be aided by the incorporation of stereo information into the same framework [3, 7].

References

[1] A. Bartoli, N. Dalal, B. Bose, and R. Horaud. From video sequences to motion panoramas. In *IEEE Workshop on Motion and Video Computing*, December 2002.

[2] S. Chien, S. Ma, and L. Chen. Efficient moving object segmentation algorithm using background registration technique. *IEEE Trans. Circuits and Systems*, 12(7):577–586, July 2002.

[3] C. Eveland, K. Konolige, and R. Bolles. Background modeling for segmentation of video-rate stereo sequences. In *Proc CVPR*, 1998.

[4] N. Friedman and S. Russell. Image segmentation in video sequences. In *13th Conf. on Uncertainty in A.I.*, 1997.

[5] X. Gao, T. Boult, F. Coetzee, and V. Ramesh. Error analysis of background adaption. In *Proc CVPR*, pages I: 503–510, 2000.

[6] D. Gutchess, M. Trajkovic, E. Cohen-Solal, D. Lyons, and A. Jain. A background model initialization algorithm for video surveillance. In *Proc. ICCV*, pages I: 733–740, 2001.

[7] M. Harville, G. Gordon, and J. Woodfill. Foreground segmentation using adaptive mixture models in color and depth. In *IEEE Workshop on Detection and Recognition of Events in Video*, 2001.

[8] T. Horprasert, D. Harwood, and L. Davis. A statistical approach for real-time robust background subtraction and shadow detection. In *Frame-Rate Workshop*, 1999.

[9] M. Irani, P. Anandan, J. Bergen, R. Kumar, and S. Hsu. Efficient representations of video sequences and their applications. *Signal Processing: Image Communication*, 8(4):327–351, May 1996.

[10] N. Jojic and B. Frey. Learning flexible sprites in video layers. In *Proc. CVPR*, pages I:199–206, 2001.

[11] A. Kitamoto. The moments of the mixel distribution and its application to statistical image classification. In *Advances in Pattern Recognition (SPR’00)*, pages 521–531, 2000.

[12] W. Long and Y. Yang. Stationary background generation: An alternative to the difference of two images. *Pattern Recognition*, 23:1351–1359, 1990.

[13] P. McLauchlan and A. Jaenicke. Image mosaicing using sequential bundle adjustments. In *Proc. BMVC*, 2000.

[14] A. Mittal and D. Huttenlocher. Scene modeling for wide area surveillance and image synthesis. In *Proc. CVPR*, 2000.

[15] A. Neri, S. Colonnese, G. Russo, and P. Talone. Automatic moving object and background separation. *Signal Processing*, 66(2):219–232, April 1998.

[16] Y. Ren, C. Chua, and Y. Ho. Statistical background modeling for non-stationary camera. *Pattern Recognition Letters*, 24(1-3):183–196, January 2003.

[17] J. Rittscher, J. Kato, S. Joga, and A. Blake. A probabilistic background model for tracking. In *Proc ECCV*,2000.

[18] S. Rowe and A. Blake. Statistical mosaics for tracking. *Image and Vision Computing*, 14(8):549–564, August 1996.

[19] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. CVPR*, pages II: 246–252, 1999.

[20] P. H. S. Torr and A. Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *CVIU*, 78:138–156, 2000.

[21] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *Proc. ICCV*, pages 255–261, 1999.

[22] Y. Wexler, A. Fitzgibbon, and A. Zisserman. Bayesian estimation of layers from multiple images. In *Proc. ECCV*, 2002.