Segmenting Video Foreground Using a Multi-Class MRF

Patrick Dickinson Andrew Hunter Kofi Appiah
University of Lincoln, UK
{pdickinson, ahunter, kappiah}@lincoln.ac.uk

Abstract

Methods of segmenting objects of interest from video data typically use a background model to represent an empty, static scene. However, dynamic processes in the background, such as moving foliage and water, can act to undermine the robustness of such methods and result in false positive object detections.

Techniques for reducing errors have been proposed, including Markov Random Field (MRF) based pixel classification schemes, and also the use of region-based models. The work we present here combines these two approaches, using a region-based background model to provide robust likelihoods for multi-class MRF pixel labelling. Our initial results show the effectiveness of our method, by comparing performance with an analogous per-pixel likelihood model.

1. Introduction

Background models are used extensively in autonomous video processing applications to detect and segment moving foreground objects. The intention is to capture a representation of an “empty” scene, so that pixel values generated by new objects can be identified by comparison. Probabilistic per-pixel models have become a de facto standard: for example, the method proposed by Stauffer and Grimson [8] uses a Mixture of Gaussians (MoG) to represent background distributions learned for each pixel.

However, robust background modelling remains only a partially solved problem. Per-pixel methods are effective under controlled conditions, but unexpected dynamic variations in the structure of the scene background result in false-positive foreground object detections. For example, such errors are typical in natural scenes characterised by moving foliage or water.

Some previous work has been directed at improving segmentation by incorporating information about the scene structure in the modelling and/or classification process. Markov Random Field (MRF) based schemes [6, 3, 5] enforce coherent labelling through local constraints, but still employ per-pixel likelihoods. Alternatively, region-based background models have also been proposed. Zhong et al. [11] used a hierarchical per-pixel method, whilst Zhang et al. [10] proposed local dependency histograms based on an $N \times N$ pixel neighbourhood.

Other methods (e.g. [4]) employ models for both the scene background and foreground objects, as a means of improving pixel classification. Our previous work [2] has shown that using spatially distributed MoGs to represent foreground and background regions separately can reduce errors caused by dynamically varying backgrounds. Recent work by Yu et al. [9] combines a similar representation with an MRF-based pixel classifier. Similarly, Sheikh and Shah [7] used non-parametric density estimates of the foreground and background likelihoods. However, the use of foreground models imposes some restrictions: prior knowledge of the foreground appearance may be required for initialisation [9] and movements between images are assumed to be small [7]. Careful parameterisation is required for online learning of parametric foreground models, as false-positive errors will be reinforced by the model.

2 Our Approach

The work that we present combines a probabilistic region-based background model with an MRF pixel classifier, and in this respect is conceptually related to the methods proposed by Sheikh and Shah [7], and Yu et al. [9]. However, unlike these, we depart from a binary foreground/background pixel labelling. This is a primary contribution of our approach.

Our premise is that each region in the background model represents a distinct process, such as part of an object, and that an observed pixel value is generated either by one such process, or by a foreground object. Each background region is used as a label for a multi-class MRF, which is then used to enforce coherent pixel
labelling at region (process) level.

We use an adaptive MoG in 5-dimensional feature space (2D image coordinates, and 3 RGB colour coordinates) to represent the background. Our model components therefore have a similar form to those used by Yu et al. [9], but we use a much larger number of components (1000) which facilitates a more detailed representation of the background.

As we have previously noted, the use of learned foreground models requires some prior assumptions about foreground appearance, dynamic behaviour, and appropriate parameterisation. We wish to avoid such restrictions, and instead use a uniform foreground likelihood. In this respect our approach is comparable with other techniques which use only a background representation, such as Migdal and Grimson’s per-pixel model[5]. Indeed, we also adapt their MRF framework (from binary to multi-class) to estimate our maximum a posteriori MAP-MRF field labelling. For this reason we use Migdal and Grimson’s method for comparison, in our evaluation in section 4.

3. Building the Background Model

The background model is learned from a single training frame, devoid of foreground objects, such that each component of the 5-dimensional MoG represents a homogeneous region of background image pixels with similar spatial and colour characteristics. Subsets of components are used to describe the local background for each pixel, so the total number of components (K) is estimated empirically such that the scene is represented in appropriate detail. An observation at pixel x, time t, is represented by a 5-dimensional feature vector \( x_{t} = [x, y, R, G, B]^{T} \), and the distribution for each model component is given by the Gaussian form:

\[
p(\mathbf{x}_{t} | \theta_{(j,t)}) = \frac{e^{-\frac{1}{2}(\mathbf{x}_{t} - \mu_{(j,t)})^{T}(\Sigma_{(j,t)})^{-1}(\mathbf{x}_{t} - \mu_{(j,t)})}}{(2\pi)^{2}|\Sigma_{(j,t)}|} \tag{1}
\]

The usual mixture model conditions hold such that each component is weighted by \( \omega_{(j,t)} \), and \( \sum_{j} \omega_{(j,t)} = 1 \). Thus the parameters \( \theta_{(j,t)} = \{ \omega_{(j,t)}, \mu_{(j,t)}, \Sigma_{(j,t)} \} \) are the weight, mean, and covariance matrix of the \( j^{th} \) component at time \( t \).

We have simplified our representation by assuming independence between the spatial and colour features. Thus the covariance matrix \( \Sigma_{(j,t)} \) has block diagonal form. This facilitates processing, as we subsequently consider the spatial and colour distributions separately. For clarity, we henceforth use the notation \( p(\mathbf{x}_{t}^{sp} | \theta_{(j,t)}^{sp}) \) when referring only to the spatial component of the distribution described by equation 1, and \( p(\mathbf{x}_{t}^{col} | \theta_{(j,t)}^{col}) \) when we refer only to the colour component.

The model parameters are estimated using the well known Expectation Maximisation (EM) algorithm [1] for MoGs. However, EM requires initial parameters estimates, which we determine using a component splitting procedure. First, a single component is created with parameters estimated from the statistics of all pixels in the training image. Its mean \( \mu_{s} \), and the principle eigenvector of its covariance matrix \( \Lambda_{p} \), are used to define a separating plane: the parameters of a second component are then estimated from those pixels satisfying: \( (x - \mu_{s}) \cdot \Lambda_{p} > 0 \). The parameters of the original component are re-estimated using the remaining pixels. The component with the highest principal eigenvalue is then selected, and split in a similar fashion. This is repeated until the target number of components have been generated, yielding an initial set of corresponding parameter estimates. The EM process is then applied for 100 iterations to ensure convergence. Figure 1 shows a visualisation of a model built for one of the sequences used in our experiments.

![Figure 1. Model Visualisation. Top: original image. Bottom: model components.](image)

For each pixel, we build a list of the \( N_{loc} \) components giving the highest value of \( p(\mathbf{x}_{t}^{sp} | \theta_{(j,t)}^{sp}) \). In our experiments we use a value of \( N_{loc} = 10 \). These lists are used to optimise the pixel classification process.
3.1 Pixel Classification

We process each image in the sequence in turn. For each image, we use the model to infer whether each pixel was most likely to have been generated by a background component, or by a foreground object. Given K background components, the corresponding possible labels are \{l_1, \ldots, l_K\}, or \{l_f\} (foreground).

Notwithstanding foreground objects, we expect that spatial variations in the background will cause the labelling of some pixels to change during the sequence. We assume that such variations are not large, and restrict the possible background labels for a given pixel \(x\) to \(\mathcal{L}_x \subset \{l_1, \ldots, l_K\}\), where \(\mathcal{L}_x\) corresponds to the pre-calculated subset of \(N_{loc}\) components giving the highest value of \(p(x|\theta^{loc}_{j,t})\). The pixel is thus labelled either as foreground (\(l_f\)), or as one of its background labels \(\mathcal{L}_x\). Restricting labelling in this way significantly reduces computational costs.

We use a multi-class MRF to enforce coherent labelling. Given the set of all possible global field labellings \(W\), segmenting an image amounts to estimating \(w \in W\) which maximises the posterior:

\[
p(w|\mathcal{I}_t) \propto p(\mathcal{I}_t|w)p(w) \tag{2}
\]

We denote the MAP value of \(w\) at time \(t\) as \(w_t\). Estimating \(w_t\) then amounts to minimising the associated energy function:

\[
E(w) = \sum_x U(x,t) (x|l(x,t)) + \sum_x \sum_{y \in \mathcal{N}_x} V(l(x,t), l(y,t)) \tag{3}
\]

where \(U(x,t)\) defines the likelihood potentials of the observed pixel values, and \(V\) defines their smoothing potentials with respect to an 8-connected neighbourhood \(\mathcal{N}_x\). Our method for estimating \(w_t\) is analogous to that used by Migdal and Grimson [5], except that we use a multi-class MRF such that \(U(x,t)\) is the exponent of \(p(x|\theta^{col}_{j,t})\) (using equation 1) if \(l_x \in \mathcal{L}_x\). For the foreground energy we use \(U(x,t) = ln 2^d\) which corresponds to the uniform distribution in 24-bit colour space. \(V\) is defined as \(-d\) if \(l_x = l_y\), or \(+d\) otherwise: in our experiments we use \(d = 3\). We estimate \(w_t\) using Gibbs sampling, employing the linear cooling schedule proposed in [5], and have experimented with between 10 and 100 field iterations to achieve convergence.

3.2 Updating the Background Model

Once learned, the spatial distributions of the background components remain invariant. However, we update the mean of the colour distributions, \(\mu^{col}_{j}\), to adapt to changes in lighting. After each segmentation we calculate the colour mean of the pixels assigned to each component, \(\mu^{col}_{j,t}\). We then update \(\mu^{col}_{j}\) using a filter constant \(\alpha\) such that:

\[
\mu^{col}_{j,t+1} = (1 - \alpha) \mu^{col}_{j,t} + \alpha \mu^{col}_{j,t} \tag{4}
\]

4 Experiments

We have evaluated our algorithm on six short video sequences. For the results presented here we used 1000 background components and set the MRF field potentials \(V\) to \(d = 3\). The number of MRF field iterations was set to 10, ranging between temperature values of 1.0 and 0.2. The learning rate \(\alpha\) was set to 0.005. For comparison, we also present results achieved using Migdal and Grimson’s scheme [5], with parameters previously estimated for this data set [2].

The sequences were filmed using a consumer DV camcorder in PAL format (720 \(\times\) 576 pixels), and were re-sampled to a frequency of 10Hz. Each sequence represents an outdoor scene with significant levels of background variation, caused by moving foliage and/or water. Each sequence also features a single foreground object: a human in five of the sequences, and a waterbird in the other. Sequences ranged from between 10 and 16 seconds in duration. This relatively short duration helped construct a meaningful ground truth. For each sequence we arbitrarily selected 16 image frames and annotated them by manually marking the pixels representing the foreground object.

We ran our algorithm on each sequence, and automatically compared the ground-truth frames with their corresponding output image segmentations. The total number of pixel classification errors was recorded for each sequence. We repeated the same process using Migdal and Grimson’s algorithm, and the results are shown in table 1. Figure 2 shows an example image from one of the sequences, with the corresponding segmentations produced by our algorithm, and Migdal and Grimson’s algorithm.

5. Discussion

We have proposed a video segmentation algorithm which combines a probabilistic region-based background model with pixel labelling using a multi-class MRF. Our results show that our approach can significantly reduce errors in scenes characterised by dynamic backgrounds, as compared with an analogous per-pixel likelihood model. We have avoided using a
Figure 2. Example Segmentation. Left to Right: Original image, Migdal and Grimson, Our Algorithm. Pixels classified as foreground are shown in white.

% Error in Pixel Classification

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Our Algorithm</th>
<th>Migdal &amp; Grimson</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.028</td>
<td>0.166</td>
</tr>
<tr>
<td>2</td>
<td>0.007</td>
<td>0.038</td>
</tr>
<tr>
<td>3</td>
<td>0.009</td>
<td>0.120</td>
</tr>
<tr>
<td>4</td>
<td>0.039</td>
<td>0.101</td>
</tr>
<tr>
<td>5</td>
<td>0.024</td>
<td>0.052</td>
</tr>
<tr>
<td>6</td>
<td>0.019</td>
<td>0.049</td>
</tr>
<tr>
<td>Mean</td>
<td>0.021</td>
<td>0.088</td>
</tr>
</tbody>
</table>

Table 1. Recorded errors by sequence

learned foreground model in order to retain generality of application. However, it is likely that for specific applications some improvement could be achieved by augmenting our background model with an appropriate foreground model. We intend to investigate this in further work. In addition, we propose to investigate more efficient methods of estimating the MAP-MRF field labelling.

6. Acknowledgements

The work presented in this paper was supported by EPSRC grant reference EP/H017143/1.

References