Background Subtraction in Dynamic Scenes with Adaptive Spatial Fusing

Zhiyu Wang \(^1\), Hui Xu \(^2\), Lifeng Sun \(^3\), Shiqiang Yang \(^4\)

\(^1\) Key Laboratory of Media and Networking, MOE-Microsoft, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua Univ. Beijing 100084, China
\(^2\) zy-wang08@mails.tsinghua.edu.cn
\(^3\) xu-h06@mails.tsinghua.edu.cn
\(^4\) sunlf@tsinghua.edu.cn
\(^5\) yangshq@tsinghua.edu.cn

Abstract—Background subtraction in highly dynamic scenes has been a critical challenge for traditional pixel-wise background models which perform poorly when the background has dynamic textures. In this paper, we consider background modelling in a spatial perspective and make an attempt to exploit more information from the outputs of pixel-wise model. We propose a background subtraction scheme using adaptive spatial fusing to refine the output of typical pixel-wise background model – Mixture of Gaussians (MoG) and employ a MRF-MAP scheme to make foreground-background classification using the spatial correlation. Experiments on several challenge sequences show that our method is able to yield significantly better results than the traditional ones and is compelling with existing state of the art background subtraction algorithms. Additionally, we proved our algorithm has linear running time complexity and any pixel-wise background model could be easily integrated into our spatial fusing scheme which greatly enhanced its scalabilities and applications.

I. INTRODUCTION

Background subtraction is a fundamental research topic in computer vision and it has been receiving attentions nowadays. A lot of approaches have been developed to segment the moving foreground from the background. Mainstream among these approaches are building and maintaining an adaptive statistical background model. When a new frame is obtained, pixels are classified as belonging to foreground or background by the likelihood of generated by the background models. For example, Stauffer and Grimson in [1] used the Mixture of Gaussians (MoG) to model pixel value distribution for background. Each pixel location consists of several Gaussian distributions with different weights in RGB or YUV or some other color spaces. They used an exponential forgetting scheme which approximates an online k-means algorithm to update each Gaussian component. Pixel which associates with uncommon Gaussian or matches no Gaussian is deemed as foreground. The matching process usually relies on a given threshold. This approach is widely applied in surveillance systems for its simplicity and real time implementation, but unfortunately, it requires not only stationary camera but also static background. When the background exhibits dynamic textures such as waving branches and leaves or waves of water, it usually mistook part of the background as moving objects.

Recently, researchers have paid more attention to exploiting outputs of MoG without employing thresholds. They try to use the data generated by comparing new coming pixel with the model to get further results with spatial, motion or color information. One of the modern approaches is to integrating the Markov Random Fields (MRF)\(^6\) into background subtraction process. The MRF model reflects the correlation of pixels in spatial and temporal neighbourhoods. To some extent, these methods alleviate the defect of pixel-wise background model in complex scenes and are less prone to the noise and background camouflaging effect than traditional pixel-wise method. Meanwhile, MRF approaches are always taken as a post-processing step, and intrinsic property of ordinary background model indeed remains unchanged. For dynamic textures, its performance is worth discussion. In our paper, we also use MRF with MoG for comparison.

In [3], Sheikh and Shah improved the kernel estimation model by combing both colour and location information of pixels. They extended Mittal and Paragios’ work\(^4\) by allowing observed pixels match kernels with nearby pixels. Small spatial motions like waving tree or water could be classified as part of background. Their work is a success and milestone of exploiting spatial correlations. Unfortunately, like most kernel based method, it must maintain a kernel history long enough which means much more computational and memory cost. Meanwhile, features of five dimensions require a high frame rate to avoid estimation failure due to sparse sample points.

Pierre-Marc introduced background model using spatial cues\(^5\). They perform background subtraction under the assumption that the spatial distribution of intensities in a neighbourhood window centred on a pixel is very similar to the temporal distributions. Thus they use spatial distribution to

\(^1\) 978-1-4244-4652-0/09/$25.00 ©2009 IEEE
approximate temporal distribution so as to reduce the calculation and memory cost. However, this assumption does not always hold, especially when there is temporal texture of different distribution with spatial information such as a blinking light.

Dalley and Grimson[6] developed an extended model of Mixture of Gaussians for each pixel of the background. They model the image generation process as arising from a mixture of components that have a Gaussian distribution in color and some spatial distribution. They use a fixed window around one centre pixel as its neighbourhood, and the value of this pixel was generated by all Gaussians from the neighbourhood locations. They use this approach to handle dynamic texture and achieve excellent results. But as the window size is hard to adjust and can hardly change adaptively, especially when the window size is quite large, the running time of comparing and updating is many times of that using a pixel-wise manner. It also brings unaffordable computational cost.

The rest of the paper is organized as follows. The model will be described in Section II, while Section III will show the detailed procedure of each part of the whole framework. Section IV explains the experiment results. We conclude this paper in Section V.

II. MODEL DESCRIPTION

As we discussed above that there are spatial correlation between pixels in a neighbourhood, we formulate our pixel generation process model as follows:

\[
p(c_i|\Phi) = f(c_i|X_{N(i)})
\]

where \(\Phi\) denotes the background model and \(c_i\) is the color information of the pixel \(i\). \(p(c_i|\Phi)\) stands for the probability of pixel \(i\) generated from the background model. \(N(i)\) refers to the set of pixels in pixel \(i\)'s neighbourhood. \(X_{N(i)}\) denotes the pixel-wise background model set for pixels belonging to \(N(i)\). This formulation provides a general format for spatial fusing based on pixel-wise background model, which can support any pixel-wise model.

In this paper, we build our background model based on the common MoG. We get adaptive neighbourhood by employing fast image segmentation. We study and make use of the unique character of pixels in the same image segment to re-formulate our general background model. After a simple and fast fusing mechanism in each segment block, we obtain a smoothed Mahalanobis distance map as the input for MRF model, and finally by constructing and minimizing an energy function, we get our foreground masks.

In this section, we will introduce the framework of our approach and present a discussion on why and how to re-formulate our general background model and get adaptive neighbourhoods for spatial fusing.

A. Framework

As shown in Figure 1, there are five parts in the framework of our background subtraction. After the input of the image frame is processed with the pixel-wise mixture of Gaussian model, the outputs are the Mahalanobis distance map from each pixel colour to its nearest MoG component. “Spatial Fusing” is the core procedure which takes into adaptive neighbourhood to perform spatial fusing on the outputs from previous steps. The refined distance map is taken as the input for the following MRF-MAP classification procedure to label pixels as foreground or background, which is finally transformed to the form of a foreground mask graph as the final result.

![Figure 1 Background subtraction framework](image)

B. Spatial Fusing Methodology

A natural idea of obtaining spatial neighbourhood generated by collective background model is to use image segmentation to cluster pixels which are near in location and appearance.

Image processing technology provides us with a lot of reliable segmentation algorithms, which greatly impel our research. We study the character of each segment produced by image segmentation, and regard each segment block as a unique logic unit in image. For any segment, it entirely belongs to either the foreground or the background, that is, all pixels in the same segment share the same foreground/background label.

If an image segment belongs to the moving foreground, given a good background model \(\Phi\) and any pixel \(i\), we will get a relatively low value of \(p(c_i|\Phi)\), and every other pixel \(j\) belongs to \(N(i)\) has a near color to \(c_i\),

\[
j \in N(i) \iff i \in N(j) \iff N(i) = N(j)
\]

According to formula (1), \(p(c_j|\Phi)\) is similar to \(p(c_i|\Phi)\). And they both fall into a relatively small range. The above discussion is based on the assumption that the segment belongs to the moving foreground, and vice versa it also holds for segments belonging to the background. As a result we can regard each pixel in the same segment sharing the same probability of generating from the background model.

Based on the observation above, we rewrite the formula (1) as following:

\[
p(c_i|\Phi) = p(N(i)|\Phi) = f'(X_{N(i)})
\]

where \(p(N(i)|\Phi)\) shows that the whole segment shares the same value, with function \(f'\) different from that of formula (1).

There are two advantages of sharing the same probability of generating from the background model. Firstly, as the pixels in the same segment share the same value, the running time and storage size can descend to a quite lower level, which makes possible the algorithm is able to process in nearly real time. Secondly, more information can be supplied to pixels so it can de-noise the background model.
III. IMPLEMENTAL DETAILS

As shown in the model, pixel-wise background model can provide sufficient information for spatial fusing. With the help of the Mahalanobis distance generated from the background model, a middle-level semantic segmentation of the image can fuse the Mahalanobis distance and fix errors of special values of some minor pixels. After that, a Markov Random Field can be built to model the object detection problem as a background/foreground classification problem and get the final background subtraction results by solving MRF-MAP functions.

A. Mahalanobis Distance

Actually, what we need from the background model is a value telling the likelihood of a pixel belonging to the background, so it is not necessary to use the probability directly. Other differentiable values can be used to reduce the calculation cost. It is well known that the MoG model can be described as:

$$P(I_t) = \sum w_i \eta(I_t, \mu_i, \Sigma_i)$$

where $w_i$ is the weight of the $i$-th Gaussian distribution and

$$\eta(I_t, \mu_i, \Sigma_i) = \frac{1}{(2\pi)^\frac{n}{2} |\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(I_t - \mu_i)^T \Sigma_i^{-1} (I_t - \mu_i)}$$

(5)

Here $\eta$ is the time label of the pixel and $I_t$ is the pixel information. $\mu_i$ and $\Sigma_i$ are the mean and variance of the Gaussian model separately, with the dimension $n$. And in formula (5), the part

$$M(I_t) = (I_t - \mu_i)^T \Sigma_i^{-1} (I_t - \mu_i)$$

(6)

is called the Mahalanobis distance. It is obvious that the larger $M(I_t)$ is, the smaller the value of $\eta(I_t, \mu_i, \Sigma_i)$ will be. So the Mahalanobis distance can be taken instead of the Gaussian probability function to show the likelihood of a pixel generated from the background model. Moreover, without the calculation of the exponential function, it is obvious to have a lower calculation cost.

B. Adaptive Neighbourhood Generation

In the extended MoG methods proposed by Dalley and Grimson[6], a fixed window is always selected to get the neighborhood information, and the running time grows with the size of window. Obviously, windows of fixed sizes can hardly adapt the dynamic scene and is fragile to different applications.

In order to perform background subtraction adaptively and to solve the running time problem, and at the same time to get a higher accuracy, we employ a fast adaptive neighbourhood with the following two advantages.

- Efficiency. The algorithm to find blocks of neighbourhood must be efficient enough (at least ten frames per second) not to be the bottleneck of the whole process.

- Adaptive Accuracy. In a higher frequency area, it should ignore tiny noise to avoid disturbing and make the neighbourhood large enough to use. On the other hand, for a lower frequency area, it should maintain small blocks out of large ones to preserve tiny objects’ properties.

According to the above requirements, Pedro’s algorithm is chosen as the segmentation algorithm[7], which produce satisfying segmentation results while be able to process tens of frames with $320 \times 240$ pixels per second and have a great adaptive accuracy over all areas of the frame.

C. Spatial Fusing

As we have already built the Mahalanobis distance image and the adaptive neighbourhood, a spatial fusing procedure can make up $f'(X_{N(i)})$. The pixel-wise background model can be regarded as a weak classifier. Now in order to get a classifier of the whole block, a simple strategy is to use the result of each pixel and do a voting. However, this method cannot make full use of the Mahalanobis distance values, and can only generate a quite rough result. On the other hand, the Mahalanobis distance values can be regarded as a new gray image with noise based on the MoG model, and spatial fusing method can make great result on that.

Median filtering is a wonderful spatial fusing function to do spatial filtering. We adopt this scheme in a different way - filtering in each segment. For each segment, the median value of Mahalanobis distance is chosen to be the value of $f'(X_{N(i)})$. The median filtering algorithm can be described as below:

$$f'(X_{N(i)}) = M(N(i))$$

(7)

where

$$M(N(i)) = \text{median}_{j \in N_i} (M(j))$$

(8)

Here $M(j)$ is the Mahalanobis distance of pixel $j$ and the function median means getting the median value over the whole neighbourhood of pixel $i$.

As stated in the above subsections, the segmentation algorithm has the advantage in running efficiency. The median filtering algorithm has an even much lower calculating complexity to accelerate the whole process. Suppose the size of a segment is $S_i$, and the total size of the image is $S$. We have

$$\sum_i S_i = S$$

(9)

It is well known that the median selection algorithm has a linear calculating complexity[8]. Therefore, for each block, the calculation complexity is $O(S_i)$, while the total complexity should be $\sum_i O(S_i) = O(\sum_i S_i) = O(S)$, which is also linear. Compared with the fixed neighbourhood approaches, adaptive neighbourhood generated by image segmentation followed by median filtering as the fusing method has a great advantage on running time. And at the same time, in the adaptive neighbourhood, nearby pixels belonging to the foreground and background separately will have a much less relations since they
belong to the different segment neighbourhood and noise of this type will be reduced significantly.

D. Markov Random Field

After updating the Mahalanobis distance with spatial fusing, a noiseless Mahalanobis distance graph has been generated. We denote $N(i)$ as $N_i$ for simple. Then Markov Random Field optimizer[8-9-10] based on $M(N_i)$ is used to classify each block as foreground or background. Our MRF minimizes the energy function below:

$$ E(\bar{i}) = E(l_i^{(t)}|l_i^{(t-1)}) = F(\bar{i}) + T(\bar{i}) + \sum_{k \in h_i} V(\bar{i}, k) $$

where

$$ F(\bar{i}) = (1 - l_i^{(t)})M(N_i) + \mu_F l_i^{(t)} $$

$$ T(\bar{i}) = \mu_T |l_i^{(t)} - l_i^{(t-1)}| $$

$$ V(\bar{i}, k) = \mu_S(\bar{i}, k)|l_i^{(t)} - l_k^{(t)}| $$

$l_i^{(t)}$ is the label of pixel $i$ in frame $t$ with 1 standing for the foreground and 0 standing for the background. $\mu_F$ is the energy of choosing a pixel to be foreground, and $F(\bar{i})$ is the total cost of deciding a foreground pixel. $\mu_T$ is the energy of applying a temporally mismatched label and $T(\bar{i})$ is the total cost of flipping. $N_i$ is the 8-connected fixed spatial neighbours of pixel $i$, and $V(\bar{i}, k)$ is the total cost of choosing a label different from a neighbour. $\mu_S(\bar{i}, k)$ is the energy of choosing a different label from a neighbour which is defined by

$$ \mu_S(\bar{i}, k) = \alpha_S \frac{\text{min}(G(N_i), G(N_k))}{\text{max}(G(N_i), G(N_k))} $$

where $G(N_i)$ is the average grey value of pixels in the segment $N_i$ and $\alpha_S$ is a const factor. It is obvious that the nearer $G(N_i)$ and $G(N_k)$ is, the larger $\mu_S(\bar{i}, k)$ can be, and $\mu_S(\bar{i}, k)$ is in the range of $[0, \alpha_S]$. 

IV. EXPERIMENTAL RESULTS

In order to show the effect and efficiency of our algorithm, we collect several videos from recent researches or public video test data. We labelled some of the frames of each video to be binary images of foreground and background as the ground truth to compare with the result of the algorithm. Specifically for the video “waving tree”, we labelled each frame which has the man in the scene as the ground truth to analyse the detailed performance of our algorithm.

A. ROC Analysis

By applying different MRF parameters, we can get different foreground/background classification results. Thus sample points of the receiver-operator characteristic (ROC) curve can be obtained by comparing them to the ground truth data we labelled by hand, which illustrate the powerfulness of proposed algorithm to traditional ones.

We conducted experiments on the “waving tree” clip and computed the true positive rate and false positive rate at frames when there is the man in the scene. With these sample points, we get a fitting curve of ROC. In Figure 2, we show the ROC curve of our result with spatial fusing comparing with that of MoG-MRF result. The blue line of our result is obviously above the red line of the MoG-MRF result. The figure shows that to get the same true positive rate, our result will have less noise than MoG-MRF, while with the same false positive rate, the spatial fusing method can have a much higher true positive rate. The ROC comparison curve shows sufficiently that our method with spatial fusing has a clear advantage over the traditional MoG-MRF background model.

![Figure 2 ROC curve for waving trees clip on various MRF thresholds](image)

B. Experiments on Different Videos

In Figure 3, we show comparisons on some frames of our test videos. The first row “Video” contains the original video scene of the specified frame. Row “Segment” is the output of the segment procedure, with each block set to the same average color of that block. Row “MoG” is a gray image standing for the scale of the Mahalanobis distance, while row “MoG final result” contains the result by using this MoG result directly to judge the foreground with MRF. Row “Spatial Fusing” is a gray image standing for the Mahalanobis distance after the spatial filtering operation on the MoG result. And row “final result” is the final result of our approach after MRF with white pixels standing for the foreground. The last row “ground truth” is our hand-labelled foreground for comparison. The fourth column “railway” has a size of 360 x 240 pixels while all others have a size of 320 x 240 or 160 x 120, so is seems a little wider when all the frames are set to the same height in the figure.

We select the “waving tree” video as our first column of Figure 3, which comes from Toyama’s paper[11] with a size of 160 x 120 pixels. The scene contains a left-right waving tree as the background, producing a dynamic environment. Then a person in blue enters the scene, which is expected to be detected by the algorithm. The fusing step makes our “final result” much better than the “MoG final result” with the same rest parameters, which has obvious false positive noise. When looking at the “MoG” Mahalanobis distance maps, noises are over the whole image, while after spatial fusing, noises become much less.

The second column called “water” comes from Zhong and Sclaroff’s work[12], whose original data has both non-object
scenes and with-object scenes, with a size of 320 × 240 pixels. The background is the sea with ocean waves which is highly dynamic, and a bottle is the expected object. We use the non-object scenes to train the background model and try to detect objects in the with-object scenes. It is clear that in “MoG” and “MoG final result”, there are large noises, while our fusing method removes these false positive errors in row “spatial fusing” and of course get a far better “final result” than that of only building a MoG model.

Mittal and Paragios’ work[4] supplied a 320 × 240 video “traffic”. They only provide even number of frames of the video such that the scene will change a lot between frames. The tree and grasses waving in the wind construct a dynamic scene. And the tree will hide cars behind it and is likely to be detected as foreground by the MoG model. In the result, MoG model will have a large block of false positive error, while our method of spatial fusing can reduce the value of Mahalanobis distance and eliminate the error with the help of MRF.

We select the fourth column “railway” from Sheikh’s research[3], which has a size of 360 × 240 pixels. There are more than 270 frames of empty scenes with the camera shaking up and down. The railroad is black and the outside is something nearly white, so it is likely to make mistakes for the shadows. A man and a car will enter the scene, heading for the opposite directions. There will be noise on the railway when the man or the car is near that and MoG model cannot deal with that directly, but ours can make a clean result with spatial fusing.

The fifth column “station” with a size of 320 × 240 pixels is an indoor video from PETS 2006 dataset[13]. Although the background is relatively static compared with the former videos, the ground is glisten where noises of shadows cannot be ignored. By comparing spatial fusing method with MoG, we find that sometimes ours can eliminate some of the shadows, which provides a much cleaner result. Additionally, the result shows the details of persons clearly, which shows the powerfulness.

We conducted our experiments on different backgrounds containing a waving tree, a water surface, a road, or some indoor scenes, and all achieved wonderful results, which shows the consistent performance of our algorithm.

C. Speed Evaluation

With a good quality of effect, our algorithm also has a perfect efficiency. In Table 1, we show the detailed running time of each procedure of the algorithm for each experimental video, including MoG, segment, spatial fusing, MRF and the total time, in the order of the table. Since the running time is short, we run the algorithm on each video for many times and get the average value of the running time to show the efficiency of the algorithm. According to the data, we can see that the spatial fusing procedure is excellent in running time, while MoG and MRF are both good performing, and the segment procedure is the most time consuming one, but is still acceptable. Actually, as described in [7], the segment algorithm can have a better performance with a better implementation. Therefore, in total, the algorithm can process about ten frames per second, which is acceptable for a video background subtraction algorithm.

Table 1 Detailed Running Time

<table>
<thead>
<tr>
<th>Video</th>
<th>Average running time per frame(in milliseconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MoG</td>
</tr>
<tr>
<td>waving tree</td>
<td>7</td>
</tr>
<tr>
<td>water</td>
<td>22</td>
</tr>
<tr>
<td>traffic</td>
<td>28</td>
</tr>
<tr>
<td>railway</td>
<td>30</td>
</tr>
<tr>
<td>station</td>
<td>26</td>
</tr>
</tbody>
</table>

At the same time, we could get more information by further analysing the running times in the table. The video “waving tree” has a size of 160 × 120 pixels and “railway” has a size of 360 × 240 pixels, while all the other three have sizes of 320 × 240 pixels. Accordingly, “waving tree” only uses about one fourth of the time of “water”, “traffic” and “station”, and “railway” uses about 9/8 of their time, which means that the algorithm has a linear calculation complexity.

V. CONCLUSION

In this paper, we studied the background modeling problem from spatial fusing perspective. By making use of the unique character of image segments, we designed a fast, robust and adaptive scheme for improving the results of traditional GMM. We achieve this by designing a simple way of smoothing the output of MoG using median filtering in image segments and we demonstrate the powerfulness of our algorithm in quality of background subtraction results, running time and drew the ROC curves of proposed method and MoG-MRF which is widely accepted as modern background modeling algorithms.

Additionally, our background scheme also provides a universal framework of other pixel-wise background model. In the future, we’ll try this algorithm on other models and develop different fusing methods to enhance the results.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under Grant No.60573167, and the National High-Tech Research and Development Plan of China (863) under Grant No.2006AA012118.

REFERENCES


Figure 3 Comparison of results