MULTIMODE SPATIOTEMPORAL BACKGROUND MODELING FOR COMPLEX SCENES

Li Sun, Quentin De Neyer and Christophe De Vleeschouwer

Institute of Information and Communication Technologies, Electronics and Applied Mathematics
Université catholique de Louvain, Belgium

ABSTRACT

We present a new approach for modeling background in complex scenes that contain motions caused e.g. by wind over water surface, in tree branches, or over the grass. The background model of each pixel is defined based on the observation of its spatial neighborhood in a recent history, and includes up to $K \geq 1$ modes, ranked in decreasing order of occurrence frequency. Foreground regions can then be detected by comparing the intensity of an observed pixel to the high frequency modes of its background model. Experiments show that our spatial-temporal background model is superior to traditional related algorithms in cases for which a pixel encounters modes that are frequent in the spatial neighborhood without being frequent enough in the actual pixel position. As an additional contribution, our paper also proposes an original assessment method, which has the advantage of avoiding the use of costly handmade ground truth sequences of foreground objects silhouettes.

Index Terms— Background subtraction, foreground detection, multiple modes

1. INTRODUCTION

Background subtraction has been investigated as a hot research topic for many years due to its wide applications in computer vision. Particularly, its output can support automatic detection and tracking of moving objects, with applications in intelligent surveillance, teleconferencing, and 3D modeling. The basic idea of background subtraction is to subtract current image from a reference model of the background, typically learnt from past observations of the scene. The subtraction leaves only non-stationary or new appearing objects.

Although it has been investigated for a long time, background modeling remains challenging, mainly due to the complexity of real natural backgrounds. In addition to illumination changes, the natural environments such as forest canopy, lawn and water surface are difficult to model, because the foreground objects blend with the background, and because the background itself changes rapidly for example due to vibrating motion patterns or to transitions between light and shadow [1]. Previous research in this area has already dealt with the problem [2–4], but few of them make full use of the spatial information around the pixel to infer plausible appearance changes due to background dynamics. Similar to [5], we believe that spatial information obtained in the neighborhood of a pixel can provide an important cue to model those changes in the appearance background pixels, especially when they are due to unpredictable motions, such as the one caused by the wind. Fig. 1 shows how the intensity value of a vegetation pixel changes between two consecutive frames, both compared to the pixel itself (blue curve) and to the closest pixel value in a 3x3 neighborhood around the pixel (red curve). The fact, that the variation of intensity defined on the pixel neighborhood is much smaller, indicate that a better prediction of the background intensity can be obtained by incorporating spatial information around pixels.

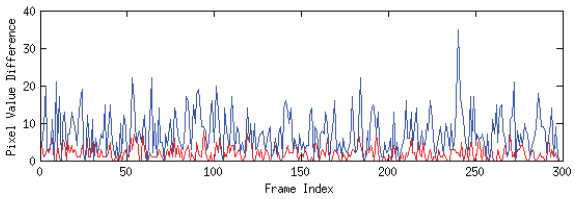

Fig. 1. Intensity variation of a single pixel as a function of time. The intensity of a pixel at a given time is compared to its intensity in the next frame (blue), or to the closest intensity observed on a 3x3 neighborhood around the pixel position, also in the next frame (red).

In this paper, we propose to model the background of a pixel not only based on its history, but also on the history of its spatial neighborhood. The two main contributions of our paper can be summarized as follows.

- First, We take advantage of both temporal and spatial information to build a background model that can preserve multiple modes that appear to be relevant in the surrounding region of the observed pixel. These modes correspond to the values that could be observed in the pixel position in case of small unpredictable movement, e.g. due to wind in grass. They also help to capture the pdf of complex backgrounds such as the ones en-

Part of this work has been funded by FP7 SV3D project, and the Belgian NSF. Thanks to ACIC for providing video sequences for evaluations.

countered on water. Specifically, modes are included to the background model of a pixel when they occur frequently enough on a spatial neighborhood around the pixel, whatever the actual position of the occurrence. This helps in capturing modes that are rare in a specific pixel, but occur frequently at random on some spatial area, e.g. like the foam on water.

- As a second contribution, our paper introduces an original assessment methodology for evaluating background models. In contrast to conventional evaluation methods, it does not require the collection of ground-truth videos, generally based on manual labeling of foreground regions. Instead, it relies on videos that do not contain any foreground object. The collection of those videos is much more easy, especially in intrusion detection contexts, which most often face empty scene.

2. RELATED WORKS

The literature on background subtraction is vast and we have limited this review to major trends. More detailed reviews can be find in [6]. Several popular methods explicitly require a off-line bootstrapping phase as a training step for learning model parameters. In this phase, the algorithm is provided with frames containing only the background. Wren *et al.* in [7] propose to model the background by a Gaussian probability density function. Stauffer *et al.* [8] instead build on a mixture of Gaussian (MOG) background model to handle the complex appearances in foreground region. Such background models lack the adaptability to the dynamic background because of the off-line training. And collecting the training frames with the strict constraint (no foreground regions) is also difficult in piratical applications.

On the other hand, some methods either do not need any off-line training [9, 10], or do not need the strict constraint on the training sample [4], and the model can automatically evolves on-line. Elgammal *et al.* [9] and Mittal *et al.* [10] propose a non-parametric method to estimate the density function for each pixel from many samples, by making use of kernel density estimation technique. Kim *et al.* [4] propose a multi-mode approach for foreground segmentation by building an efficient codebook (CB) model, which is designed to capture the variation in the background. Although the method can deal with periodic variations over time, it ignores the spatial information around the pixel, and is not able to capture complex spatio-temporal distributions of background values such as the one encountered on water. To mitigate the impact of motion on the background model learning rate, Barnich and Droogenbroeck [5] propose to learn and update the background model of a pixel based on a pseudo-random sampling of the pixel neighborhood. Their method considers a mode is relevant for background, because the mode frequently appears in either the temporal domain or the spatial neighborhood. But the success of previous works, which only use the temporal information, shows that repetition in temporality should be decoupled from the spatio-temporal domain, and regarded to be more important. Our method instead rely on the observation of the entire pixel neighborhood to update a model composed of several modes, characterized by their mean-value and a metric reflecting their frequency of occurrence. In addition, we the carefully control the way a mode is incorporated or rejected from the set of so called active modes, which are the ones that actually characterize the background appearance.

3. PROPOSED ALGORITHM

In order to be successful in real applications, background subtraction techniques have to deal with following considerations: (1) how to build the model? (2) how to update model over time? and (3) how to classify a given pixel as foreground or background, based on the model? In this section, we give answers to these questions in detail.

3.1. Background Model Format

The proposed multi-mode spatio-temporal model $\mathrm{M}_t(i,j)$, built for each pixel $\mathbf{x}_t(i,j)$ sampled at the coordinate position (i,j) from the image at time t, is composed of several modes $\mathrm{M}_t(i,j) = \{\mathbf{m}_0, \mathbf{m}_1, \cdots, \mathbf{m}_{K-1}\}$. Here K is a constant the total number of existing modes, and $\mathbf{m}_k$ represents the k^{th} mode when modes are ordered in decreasing order of occurrence. Each mode $\mathbf{m}_k$ corresponds to a value that is observed frequently enough in the neighborhood of the pixel (i,j). It consists of the color vector $\mathbf{v}_k = (Y_k, U_k, V_k)$ indicating the color center of the mode in YUV color space, and the occurrence metric f_k , which is computed as explained in Section 3.2 to reflect the occurrence frequency of mode $\mathbf{m}_k$. f_k is set to zero to indicate the mode $\mathbf{m}_k$ is still vacant. Note that the color vector can be represented either in RGB color space or YUV color space. In this paper, we choose the YUV color space because researchers find that it is easier to distinguish between shadows and objects in this color space [11]. $\mathrm{M}_t(i,j)$ summarizes the information related to the observations made in a recent history of the spatial neighborhood of pixel (i,j), as detailed in the next section.

3.2. Model Update and Foreground Detection

A pseudo-code that summarizes the textual description that follows is given at the end of the section.

3.2.1. Model Update

In this section, we explain on how to update the background model continuously based on new frames observations after its initialization, which can be simply performed by setting all the domains in the first ranking mode $\mathbf{m}_0$ as following $\mathbf{v}_0 \leftarrow \mathbf{x}_t(i,j); f_0 \leftarrow 1$.

The classical approaches to update a background model replace or down-weight old observations compared to recent ones [8]. In this paper, we adopt a similar strategy to monitor the multiple modes $\mathbf{m}_k$ in the model $\mathrm{M}_t(i,j)$ for each pixel. Formally, the modes $\mathbf{m}_k$ is defined based on a spatial covering region with its center located at every pixel position in the image. Only those pixels lying in the covered region are considered to control the the model update. In other words, the observation $\mathbf{x}_t(m,n)$ is taken into the model $\mathrm{M}_t(i,j)$ if position coordinates m and n satisfy following equations.

$$\begin{aligned} i - \lfloor N/2 \rfloor \leq m \leq i + \lfloor N/2 \rfloor \\ j - \lfloor N/2 \rfloor \leq n \leq j + \lfloor N/2 \rfloor \end{aligned} \tag{1}$$

Here N is an odd number chosen from $1, 3, 5, \cdots$, and $\lfloor \rfloor$ is the symbol of the floor function. Let $\mathbf{x}_t(m,n) = (Y, U, V)$ denote the color vector observed in the $N \times N$ neighborhood surrounding pixel at coordinate (i,j), as defined in Equation (1). For each mode $\mathbf{m}_k$ in $M_t(i,j)$, the distances $d_k(m,n)$ between the existing mode center $\mathbf{v}_k$ in $\mathrm{M}_t(i,j)$ and a given observation $\mathbf{x}_t(m,n)$ are calculated as:

$$d_k(m,n) = d_L + \alpha d_C \tag{2}$$

where $d_L = |Y - Y_k|$ and $d_C = |U - U_k| + |V - V_k|$ respectively correspond to luminance and chrominance distances. [1] For the observation $\mathbf{x}_t(m,n)$, we define $k_{\min}(m,n)$ as following:

$$k_{\min}(m,n) = \arg\min_k d_k(m,n) \tag{3}$$

In the case that $d_{k_{\min}}(m,n)$ is larger than T_{dis}, a new mode should be created because no existing modes are similar to $\mathbf{x}_t(m,n)$. In practice, when $K = K_{\max}$, the mode is simply created by replacing the mode with smallest occurrence metric. In contrast, if $d_{k_{\min}}(m,n)$ is smaller than a threshold T_{dis}, the occurrence metric $f_{k_{\min}}$ of $\mathbf{m}_{k_{\min}}$ is increased by a fixed number fi, unless it reaches the maximum saturation value fm or it has already been incremented at time t. To understand why the occurrence metric f_k of a mode $\mathbf{m}_k$ is only incremented once at every time instant, independently of the possible repetitions of the mode in the $N \times N$ neighborhood, it is useful to refer to Section 3.2.3, and anticipate the definition of an active mode. Formally, a mode $\mathbf{m}_k$ is defined to be active when its f_k is above some threshold. In that case, the mode $\mathbf{m}_k$ is considered to be frequent enough to describe the appearance of a background object, and will thus be considered to estimate the background mask. Hence, limiting the magnitude of occurrence frequency increments at each time instant allows to control the update rate of the model. As explained in Section 2, this update rate is a critical parameter of any background model adaptation strategy since it fundamentally trades off the adaptation to background changes (lack of adaptiveness leads to false detections) and the inclusion of foreground objects in the background model (leading to missed detections). In the case which $d_{k_{\min}}(m,n) < T_{dis}$, the color center $\mathbf{v}_{k_{\min}}$ of $\mathbf{m}_{k_{\min}}$ also needs to be updated. Here we adopt an exponentially weighted moving average as defined by (4). β is a parameter, controlling the update rate for $\mathbf{v}_{k_{\min}}$, whose value is in the range $(0,1)$.

$$\mathbf{v}_{k_{\min}} \leftarrow (1-\beta) \cdot \mathbf{v}_{k_{\min}} + \beta \cdot \mathbf{x}_t(m,n) \tag{4}$$

After all the observations $\mathbf{x}_t(m,n)$ in the spatial neighborhood have been considered, the f_k for all the modes are decreased by a fixed number fd. Decreasing f_k gives the opportunity to lower the ranking position for the modes that did not appear anymore in a recent history, thereby allowing for background model updates. Finally all modes $\mathbf{m}_k$ are ranked in decreasing order of their updated occurrence metrics f_k.

3.2.2. Foreground Detection

Given the background model, activated modes can be defined by measuring the occurrence metric f_k. If $f_k \geq T_f$, $\mathbf{m}_k$ is considered to be activated. Foreground regions can then be directly segmented from image by computing the distances d_k between the sample $\mathbf{x}_t(i,j)$ and the centers $\mathbf{v}_k$ of all activated modes $\mathbf{m}_k$ as is shown in (2). Then we find the smallest distance $d_{k_{\min}}(i,j)$ among all d_k. Only if the minimum distance $d_{k_{\min}}(i,j)$ is larger than a threshold T_{dis}, the pixel is assumed to be in the foreground region, and the binary foreground mask $mask_t(i,j)$ is set to 1; otherwise, it is assigned to the background and $mask_t(i,j)$ is set to 0.

3.2.3. Impact of Parameters on Model Update Rate

This section aims at highlighting the fact that the parameters fi, fd, fm and T_f, involved in model update and foreground mask computation, are crucial because they actually determine the adaptation speed that the background model reacts and adapts to dynamic backgrounds. As explained in Section 2, the update strategy is a critical component of most background models, because it trades off between the false foreground detections caused by a too slow adaptation to background changes, and the missed detections due to the inclusion of a foreground object into the background model.

With that respect, the ratio between fi and fd appears to be an important factor since it controls the threshold beyond which a mode will be considered as being frequent enough to be relevant. T_f is an important threshold, which prevents the immediate inclusion of foreground objects in the background model. fm actually controls the possible longest time that a mode is kept in the background model since its last occurrence. Actually, fi, fd, and T_f together define the background model update latency, which is important regarding

[1] Without loss of generality, more complex distance metrics can also be adopted, e.g. as described in [4]. The value range of α in (2) is from 1.0 to 1.5 and we set its value to 1.2. A large weight is given to the chrominance component because it is more stable than luminance if there are illumination changes in the environment.

the inclusion of new relevant modes in the background model. Large value for T_f and small value for fi, fd result in the high model latency but stronger robustness against inclusion of foreground objects into the background model. In practice fd is set to one by default, and fi is selected to be larger than one, typically 2 or 3. T_f and fm are set to 200 and 2000 respectively. Although so many parameters are difficult to tune manually, they also give the opportunity to be adapted in different deployments, which makes the method more flexible in some degree.

Algorithm 1 Proposed algorithm for background subtraction

Input: input image $\mathbf{x}_t$, image height and width h, w, allowable modes number K, and size of neighborhood N
Output: binary image $mask_t$ and background model $\mathbf{m}_k^t = \{\mathbf{v}_k^t, f_k^t, flg_k^t\}$

```
Initialization: t = 0;
for i < w and j < h do
    v_0 ← x_t(i, j); f_0 = 1; flg_0 = 1;
end for
Model update and foreground detection:
while t ≠ end do
    for i < w and j < h do
        Model update
        for m and n satisfied equation (1) do
            k_min = arg min_k d_k(m, n) with d_k(m, n) defined in (2);
            if d_{k_min} ≥ T_dis then
                v_{K-1} ← x_t(m, n);
                f_{K-1} ← f_{K-1} + fi; flg_{K-1} = 1
            else if flg_{k_min} = 1 then
                flg_{k_min} = 0;
                v_{k_min} ← (1 − β) · v_{k_min} + β · x_t(m, n)
                if f_{k_min} < fm then
                    f_{k_min} ← f_{k_min} + fi;
                end if
            end if
        end for
        Foreground detection
        k_min = arg min_k d_k(i, j) with d_k(i, j) defined in (2);
        if d_{k_min} < T_dis and f_{k_min} > T_f then
            mask_t(i, j) = 0;
        else
            mask_t(i, j) = 1
        end if
        for k < K do
            f_k ← f_k − fd; flg_k = 1;
        end for
    end for
end while
```

4. ASSESSMENT METHODOLOGY

This section proposes an original assessment methodology for quantitative evaluation of the result of background subtraction. Our purpose is to avoid referencing to sequences with manually labeled foreground regions, mainly because the generation of such handmade sequences is a heavy task that can not be implemented each time a system is deployed in a real surveillance environment. Hence, we are interested in assessment metrics that can be easily and automatically computed based on sequences without foreground objects, because those sequences are easily available in many practical scenarios, including in intrusion detection surveillance systems for example. For such sequences, all pixels that are labeled as foreground by the background subtraction algorithm can be considered as false alarms, which allows direct computation of the false positive rate Fp.

Another metric Bd, which actually reflects the false negatives (or missed detections), is defined by summing up the subspace covered by the multiple modes of the background model in the investigated color space. If the background density Bd is larger, the background detection can resist to a bigger noise, which results in a lower Fp, but it also increases the risk of possible missed detection. So we need to consider the trade off. Here, both Bd and Fp are obtained by taking the average on all image frames in the sequence.

Typically, Bd can be controlled by the threshold T_{dis}. By gradually increasing T_{dis}, we can obtain a series of Fp values. In other words, several pairs of coordinates (Bd, Fp) can be obtained by choosing the different T_{dis} and we can draw a Relative Operating Characteristics (ROC) based on (Bd, Fp) as is shown in Fig.2. In this curve, points near the origin indicate better performance compared to those far away from it. The crossover points between curves reveal that a method can be better than another within some range of Bd, while performing worse in other Bd value range.

The proposed assessment method can be used either to set the value of some critical parameters described in previous section such as the threshold T_{dis} or the size of the neighborhood N, or to compare performances of two different background subtraction methods. However, a fair comparison of different background subtraction algorithms can only be provided if these methods have a similar background model update latency, which is the average time needed by a model update mechanism to account for the change of background appearance. However, the latency is controlled by a set of parameters that depend on the details of each method. In practice, they should be set to lead to comparable latency so that it makes the comparison of different methods relevant.In the next section, we give explanations on the experimental result based on the proposed assessment scheme.

5. EXPERIMENTS

To evaluate the result of the proposed method, we have run experiments on several sequences characterized by complex dynamic backgrounds. There are swaying trees, grass and floating water surfaces in these sequences.

5.1. Result Analysis

Fig.2 analyzes the impact of the size of neighborhood N. N is set to $1, 3, 5$ and several pairs of coordinates (Bd, Fp) are generated by using different thresholds T_{dis}. Here, $N = 1$ implies that no spatial information is used in the background model. We observe in Fig. 2 that the best value for N depends on the content of the sequence, and on the targeted Fp. If there is large motion in the background, a large value of N is useful to decrease Fp. As is shown on the left of Fig.2, $N = 5$ gives the best performance in the range of small Fp values. The reason is that the sequence "Water Surface" contains a background with large irregular movement on water surface. In contrast, $N = 3$ leads to the best performance on the right side of Fig.2, mainly because the motions of the background

caused by trees and grass oscillation are small. Extensive use of our method reveals that for most sequences, $N = 3$ and $T_{dis} = 20$ gives a good balance between Bd and Fp.

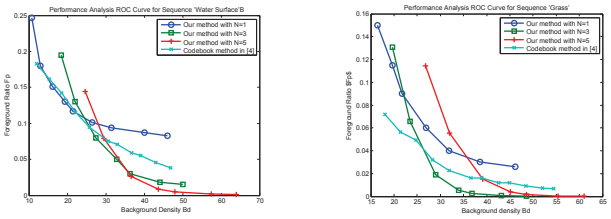

Fig. 2. Performance comparison between our method, with different values of N, and the method in [4].

Fig.2 also presents a comparison with the codebook algorithm in [4], which is a typical multi-mode background subtraction method based on temporal historical observations of a single pixel. In contrast to our approach, [4] has the opportunity to analyze the more image frames as the training dataset before defining the background model. Hence it has the opportunity to exploit the future, and is not appropriate when dynamic background model adaptation is required. In contrast, our approach can adapt to changing backgrounds. However, to make the comparison relevant, we have set the parameters fi, fd, fm, T_f and β to values[2] that promote slowness of background model adaptation, thereby avoiding that rapid (compared to the background change rate) update of the model helps in reducing the Fp rate. However, we observe that our proposed approach, whilst more flexible and adaptive, achieves better performance than [4] when $N = 3$ or 5.

5.2. Visual Comparison

In Fig. 3, several frames are presented from different sequences. The foreground detection results obtained with [4], [5], and our proposed method are shown in the 2nd, 3rd and 4th columns respectively. In the 1st and 2nd row in Fig. 3 , there is no actual foreground region and our method generates less noises than the other two schemes. In the 3rd row, there are two persons walking in a complex scene. Our method generates less noise while still detecting the two persons' silhouettes. Note that we did not tune parameters for each sequence. Instead, we apply the default parameter in all methods.

6. CONCLUSION

In this paper, a multi-mode spatio-temporal background modeling algorithm is proposed for detecting foreground regions in complex scenes. Multiple modes in the model can reflect the possible temporal variations for pixels that lie in the background region. Besides, the analysis of the pixel appearance history in a spatial neighborhood helps in capturing the set of modes associated to moving backgrounds (e.g. wind in grass) or to complex but stationary in time and space stochastic distributions of background values, as encountered on water surfaces (foam). Comparison with previous arts shows that the proposed method has better performance on the precision and recall ratio for foreground region in the image.

[2] $fi = 2$, $fd = 1$, $fm = 2000$, $T_f = 200$ and $\beta = 0.01$

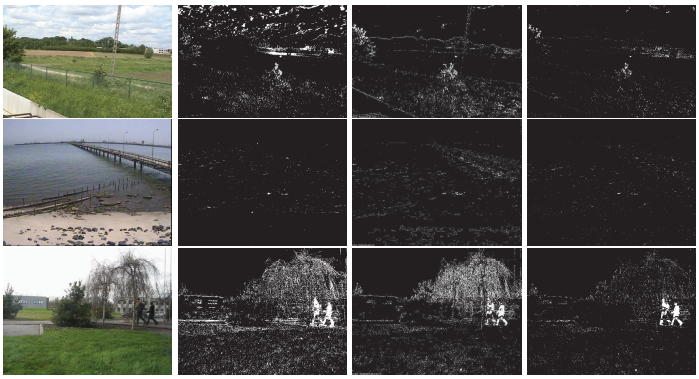

Fig. 3. Selected foreground detection results. The original image, and the results obtained with [4], [5] and the proposed algorithm are shown in each row respectively.

References

[1] T. Ko, S. Soatto, and D. Estrin, "Background subtraction on distributions," *Computer Vision–ECCV*, pp. 276–289, 2008.

[2] T. Ko, S. Soatto, and D. Estrin, "Warping background subtraction," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2010, pp. 1331–1338.

[3] P. Noriega and O. Bernier, "Real time illumination invariant background subtraction using local kernel histograms," *British Machine Vision Association (BMVC)*, pp. 567–580, 2006.

[4] K. Kim, T.H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground-background segmentation using codebook model," *Real-time imaging*, vol. 11, no. 3, pp. 172–185, 2005.

[5] O. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," *Image Processing, IEEE Transactions on*, , no. 6, pp. 1709–1724, 2011.

[6] M. Cristani et al., "Background subtraction for automated multisensor surveillance: a comprehensive review," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 43, 2010.

[7] C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, "Pfinder: Real-time tracking of the human body," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 780–785, 1997.

[8] C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 1999, vol. 5, pp. 246–252.

[9] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," *Computer Vision–ECCV*, pp. 751–767, 2000.

[10] A. Mittal and N. Paragios, "Motion-based background subtraction using adaptive kernel density estimation," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2004, vol. 2, pp. II–302.

[11] N. Martel-Brisson etal, "Moving cast shadow detection from a gaussian mixture shadow model," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2005, vol. 2, pp. 643–648.