VIBE: A POWERFUL RANDOM TECHNIQUE TO ESTIMATE THE BACKGROUND IN VIDEO SEQUENCES

Olivier Barnich and Marc Van Droogenbroeck

University of Liège
Montefiore Institute, INTELSIG Group
Liège, Belgium

ABSTRACT

Background subtraction is a crucial step in many automatic video content analysis applications. While numerous acceptable techniques have been proposed so far for background extraction, there is still a need to produce more efficient algorithms in terms of adaptability to multiple environments, noise resilience, and computation efficiency. In this paper, we present a powerful method for background extraction that improves in accuracy and reduces the computational load. The main innovation concerns the use of a random policy to select values to build a samples-based estimation of the background. To our knowledge, it is the first time that a random aggregation is used in the field of background extraction. In addition we propose a novel policy that propagates information between neighboring pixels of an image. Experiment detailed in this paper show how our method improves on other widely used techniques, and how it outperforms these techniques for noisy images.

Index Terms— Surveillance, Pattern recognition, Signal analysis, Video signal processing

1. INTRODUCTION

Background subtraction is one of the most widely used tool in automatic video content analysis, especially in video-surveillance. Numerous methods for background subtraction techniques have been proposed over the years (see [1, 2] for surveys). In most of them, a model of the recent history is built for each pixel location. The classification of new pixel values is achieved by comparing each of them to the corresponding pixel models. These techniques can be divided in two categories: (i) *parametric techniques* that use a parametric model for each pixel location and (ii) *samples-based techniques* that build their model by aggregating previously observed values for each pixel location.

The Gaussian Mixture Model [3] is probably the most popular parametric technique. It is adaptive and able to deal with the multi-modal appearance of the background of a dynamic environment (changing time of day, clouds, tree leafs,...). However since its sensitivity cannot be accurately tuned, its ability to successfully handle high- and low-frequency changes in the background is debatable, as detailed in [4]. Furthermore the estimation of the parameters of the model (especially the variance) can become problematic for noisy images.

Samples-based techniques [4, 5, 6] circumvent a part of the parameters estimation step by building their models from observed pixel values. This enhances their robustness to noise. They provide fast responses to high-frequency events in the background by directly including newly observed values in their pixel models. However, their ability to successfully handle concomitant events evolving at various speeds is limited since they update their pixel models in a first-in first-out manner. As a matter of fact, some of them use two sub-models for each pixel [4, 5]: a short term model and a long term model. While this can be a convenient solution, it is artificial and requires fine tuning to work properly in for any given situation.

This paper presents a samples-based algorithm for background subtraction. The first contribution is a novel a random selection policy that ensures a smooth exponentially decaying lifespan for the sample values that constitute the pixel models. It makes it able to successfully deal with concomitant events with a single model of a reasonable size for each pixel. The second contribution is related to the post-processing on which all the abovementioned methods rely to give some degree of spatial consistency to their results. For that purpose, we use an innovative, fast, and simple spatial information propagation method that randomly diffuses pixel values across neighboring pixels. Accordingly, our method is able to produce spatially coherent results directly. As a third contribution, we provide and instantaneous initialization technique that makes our algorithm usable starting from the second frame of a sequence.

Section 2 describes our new background subtraction algorithm. Experimental results are detailed in Section 3. Section 4 concludes the paper.

The first author has a grant funded by the FRIA (Belgium).

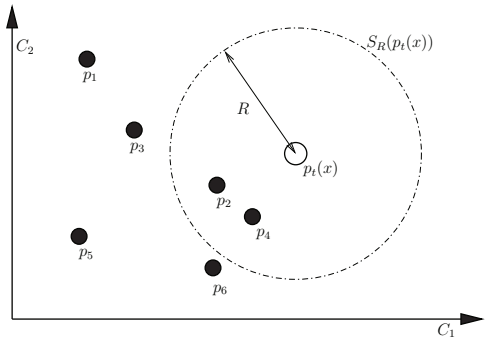

Fig. 1. To classify $p_t(x)$, we count the number of samples contained in the sphere of radius R around $p_t(x)$.

2. VIBE: A NEW ALGORITHM FOR BACKGROUND SUBTRACTION

We now describe our background subtraction algorithm. We call it ViBe, which stands for "Visual Background Extractor". We begin by defining the pixel model we use to estimate the background.

2.1. Pixel model and classification process

We denote by $p_t(x)$ the value at time t of the pixel x. Many advanced techniques (including kernel density estimation [4] and the gaussian mixture model [3, 7] are used to provide an estimate of the temporal probability density function (pdf) of a pixel x. Once the model is built, the algorithm will classify a pixel value $p_t(x)$ as a background or foreground pixel value depending on how it "fits" within the estimated pdf. A major drawback of this approach is that the evaluation of the pdf is a global process; outlayers stored in the pixel model will change the shape of the pdf although their values in the polychromatic space might be distant to $p_t(x)$.

Our approach is considerably different. We impose the influence of a value in the polychromatic space to be limited to the local neighborhood. In practice, we do not estimate the pdf, but use a set of sample values as a pixel model. To classify a value $p_t(x)$, we compare it to its *closest* values among the set of samples by defining a sphere $S_R(p_t(x))$ of radius R centered on $p_t(x)$. A pixel value is then classified as background if the cardinality, denoted $\sharp$, of the set intersection of this sphere and the set of samples $\{p_1, p_2, \ldots, p_n\}$ (see Figure 1) is above a given threshold $\sharp_{\min}$. More formally, we compare $\sharp_{\min}$ to

$$\sharp\{S_R(p_t(x)) \cap \{p_1, p_2, \ldots, p_n\}\}. \quad (1)$$

Two parameters determining the accuracy of our model are the radius R of the sphere and the minimal cardinality $\sharp_{\min}$. Experiments have shown that a unique radius R and a cardinality of 2 offers excellent performances. There is no need to adapt these parameters during the background subtraction nor do we need to change them for different pixel locations within the image. Note that since the number of samples n and $\sharp_{\min}$ are chosen to be fixed and impact on the same decision, the sensitivity of model can be adjusted using the $\frac{\sharp}{n}$ ratio.

2.2. Model update over time

To achieve accurate results over time and to handle new objects that appear in a scene, the model has to be updated regularly. Since with our model we compare x_t directly to the samples, the question on which samples have to be kept by the model and for how long is of crucial importance. In Section 2.2.1, we propose an original lifespan policy for the values over time.

The question of including or not foreground pixels values in the model is always raised when designing a background estimation method. Conservative update (no inclusion of foreground values in the model) seems, at first, to be the obvious choice but leads to deadlock situations if background objects suddenly start to move (*e.g.* parked car). On the contrary, blind update is likely to include foreground information in the background model when encountering slow moving targets. Rigorously speaking, temporal information is not available when the background is masked. As background subtraction is a spatio-temporal process, the best fallback strategy consists to exploit spatial information. This is the role played by the information propagation method proposed in Section 2.2.2. Consequently, we can afford the use of a conservative update scheme. Ghosts caused by moving background objects gradually disappear as information about the background evolution diffuses from the neighboring pixels.

2.2.1. Sample values lifespan policy

Previous methods use first-in first-out policies to update their models. To properly deal with wide ranges of events in the scene background, some authors (*e.g.* [6]) choose to include large numbers (up to 200) of samples in the pixels models. Others [4, 5] even incorporate two temporal sub-models to successfully handle fast and slow modifications.

From a theoretical point of view, it might be better to guarantee a monotonic decay for the probability of a sample value to remain inside the set of samples. It improves the relevance of the estimation and allows the use of fewer samples. We manage to do this by choosing, randomly, which sample to replace when updating a pixel model. Once the sample to be discarded has been chosen, the new value replaces the discarded sample (see Figure 2). Mathematically, one shows that, according to this updating mechanism, the probability for a pixel sample present a time t_0 to be still present at a later time t_1 is $\left(\frac{n-1}{n}\right)^{t_1-t_0}$, which can be rewritten as

$$P(t_0, t_1) = e^{-\ln(\frac{n}{n-1})(t_1-t_0)}. \quad (2)$$

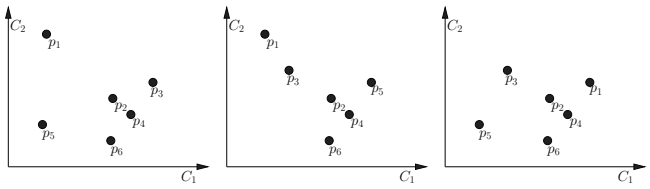

Fig. 2. This figure shows 3 of the n equally probable possible models after the update of the model shown on Figure 1.

2.2.2. Spatial consistency

Previous explanations and techniques do ignore the pixel location inside the image. But to ensure the spatial consistency of the whole image model and handle practical situations such as small camera movements or slowly evolving background objects, we adopt a technique similar to that developed for the updating process in which we choose at random and update a pixel model in the neighborhood of the current pixel. Let us denote by $N_G(x)$ and $p(x)$ respectively the spatial neighborhood of a pixel x and its value. Assume that it was decided to update the set of samples of x by inserting $p(x)$. Then we also use this value $p(x)$ to update the set of samples of one of the pixels in the neighborhood $N_G(x)$, chosen at random.

Since pixel models contain many samples, irrelevant information that could accidentally be inserted into the neighborhood model do not affect the accuracy of the detection. Furthermore, the erroneous diffusion of irrelevant information is blocked by the need to match an observed value before it can further propagate. This natural limitation inhibits the diffusion of error. With the operational constraints of very low noise levels and a fixed camera, our approach to ensure spatial consistency presents results similar to other techniques. However when these constraints are not met, for example when the camera moves or in the case of low illumination, ViBe outperforms other techniques. In Section 3, we will further analyze the segmentation results in the presence of noise.

Note that none of the selection policy or the spatial propagation is deterministic. In other words, if the algorithm is run over the same image again, the results will always differ. Although unusual, the strategy to let a random process decide on the samples to be discarded proves to be very powerful. This is different from known strategies that introduce a fading factor or that uses a long term and a short term history of values.

3. EXPERIMENTAL RESULTS

All our tests were conducted using a model containing 20 sample values, a sphere of radius 30 and a cardinality of 2 (see equation 1). The set of parameters is fixed (there was no parameters tuning).

3.1. Model initialization

Although the model could easily recover from any type of initialization, for example by choosing a set of random values, it is convenient to get an accurate background estimate as soon as possible. Ideally we would like to be able to segment the video sequences starting from the second frame, the first frame being used to initialize the model. Since no temporal information is available prior to the second frame, we populate the pixel models with values found in the spatial neighborhood of each pixel. More precisely, we fill them with values randomly taken in their neighborhood on the first frame. This strategy proved to be successful: the background estimate is valid from the second frame. The only drawback is that the presence of a moving object in the first frame will introduce a ghost object that has to fade over time.

3.2. Comparison with the gaussian mixture model

The Enhanced Gaussian Mixture Model (EGMM) presented in [5, 7] is representative as a state-of-the-art deterministic parametric method for background subtraction. A visual comparison of the results obtained using it and ViBe leads to the conclusion that, while being different, both techniques give overall equivalent segmentation accuracy for indoor sequences. The higher false positives rate of the EGMM algorithm is balanced by the higher false negatives rate of our algorithm. However, foreground objects boundaries are sharper with our technique.

Outdoor cameras sequences lead to different observations as images are often of poorer quality. They suffer from camera shake caused by the wind, higher level of background motion and high image compression for bandwidth reduction issues. One of our test sequences was taken on a rainy and windy day with a strong compression ratio. While ViBe managed to keep honorable results, the EGMM algorithm clearly suffered more from these difficult conditions. The noise seems to prevent the EGMM algorithm from estimating correctly the parameters of its model.

3.2.1. Robustness against noise

To be able to objectively compare EGMM and ViBe in noisy conditions, we produced ground truth segmentation maps of an outdoor sequence using a blue-screen like process. We added salt-and-pepper noise to the ground-truth sequence, and computed the precision and the recall for several levels of noise.

Noisy images were produced by adding a uniformly distributed random noise on all the pixels. The noise was introduced to produce PSNR's ranging from 16 [dB] to 51 [dB].

The visual results shown in Figure 3 are self-explanatory. Nevertheless, let's look at the precision and recall curves of Figure 4. The curves clearly exhibit that ViBe resists to important additive noise levels, even for PSNR's as low as

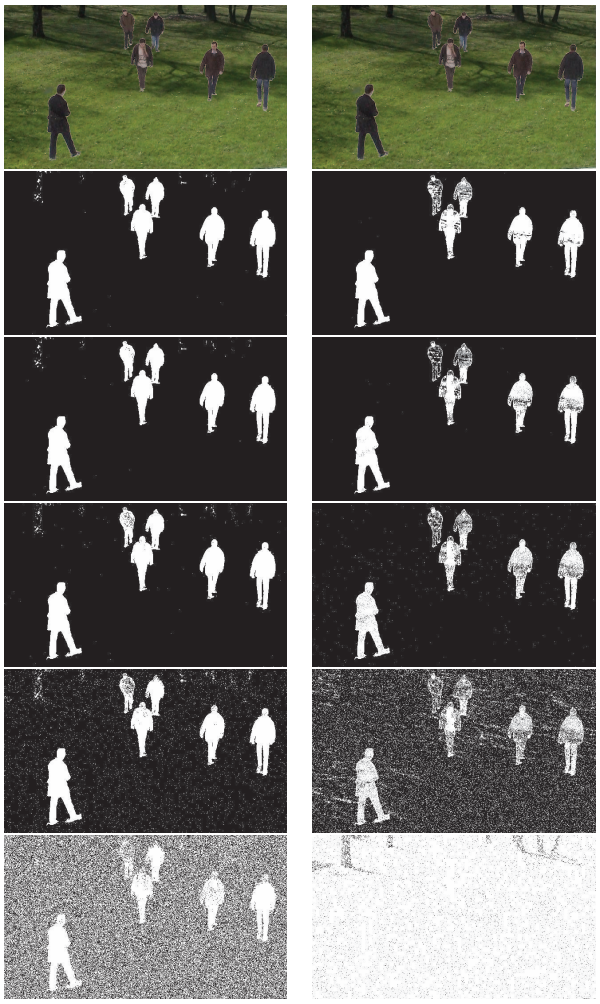

Fig. 3. Results of ViBe (left) and of EGMM (right) for PSNR's of 51 [dB], 39 [dB], 30 [dB], 25 [dB], and 19 [dB].

30 [dB]. On the other side, the precision of the EGMM algorithm decays very quickly, even for small amounts of added noise. In areas where precision remains at meaningful levels, the recall curve of ViBe in RGB color-space lays in between the one obtained by our technique in the HSV colorspace and the one obtained by the EGMM algorithm.

4. CONCLUSIONS

In this paper, we present a novel samples-based background subtraction algorithm called ViBe. Thanks to our update policy, our spatial information propagation method, and our instantaneous initialization technique, we are able to deal with several concomitant events evolving at various speeds, to get spatially coherent segmentation maps without any form of post-processing, and to get meaningful results starting from the second frame of a sequence. Furthermore, since ViBe can afford the use of a strictly conservative update scheme and directly compares the pixel values to the samples stored in the pixel models, it exhibits an excellent robustness to noise.

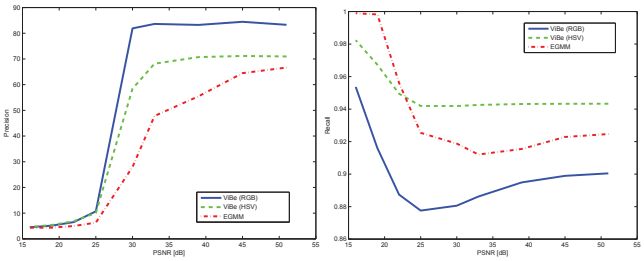

Fig. 4. Precision (left) and recall (right) curves of ViBe and the EGMM algorithm for several PSNR's.

During our experiments, we used ground truth segmentation maps of an outdoor sequence to compare the results of ViBe with those of an independent state-of-the-art technique (EGMM presented in [7]). We showed that while the performances of EGMM decay very quickly in the presence of noisy data, our technique strongly resists to important amounts of noise. It even manages to produce accurate results for PSNR's as low as 30 [dB]. ViBe does not require any fine tuning; it produces good results in various environments with a fixed set of three parameter values (number of samples stored for each pixel model, matching threshold between a pixel value and a sample, and number of matches needed to incorporate a pixel into the background of the scene).

5. REFERENCES

[1] M. Piccardi, "Background subtraction techniques: a review," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, 2004, vol. 4, pp. 3099–3104.

[2] R. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: A systematic survey," *IEEE transactions on image processing*, vol. 14, no. 3, pp. 294–307, 3 2005.

[3] C. Stauffer and E. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, 2000.

[4] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," in *ECCV '00: Proceedings of the 6th European Conference on Computer Vision-Part II*, London, UK, 2000, pp. 751–767, Springer-Verlag.

[5] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773–780, 2006.

[6] H. Wang and D. Suter, "A consensus-based method for tracking: Modelling background scenario and foreground appearance," *Pattern Recognition*, vol. 40, no. 3, pp. 1091–1105, 2007.

[7] Z. Zivkovic, "Improved adaptive gausian mixture model for background subtraction," in *Proceedings of the International Conference on Pattern Recognition*, 2004, pp. 28–31.