MVA2002 IAPR Workshop on Machine Vision Applications, Dec. 11 - 13, 2002, Nara - ken New Public Hall, Nara, Japan

15—1

A Novel Background Initialization Method in Visual Surveillance

Alessandro Bevilacqua, Member, IEEE*
ARCES - DEIS (Department of Electronics, Computer Science and Systems)
University of Bologna

Abstract

The background subtraction is a common method for real-time segmentation of moving targets in image sequences. This could be a true image without moving objects. However, often a background free of moving objects is not available, therefore a model should be employed. Most of the research works dealing with a background model cope with its updating, but not with its initialization. In this paper we propose an original method which is able to effectively extract a reliable stationary background having at disposal a short sequence with an undetermined number of foreground objects. It is based on the improving of a likelihood-based background model by using information about reliable stationary pixels achieved through a simple motion detection algorithm.

1 Introduction

A monitoring system, as well as a visual surveillance system, initiates targets identification by determining which parts of each image in a sequence belong to moving objects and which to the background. Background differencing is an effective technique apt to detect moving pixels when sequences come from a stationary camera. A common way to accomplish this step consists of examining the difference in pixel intensities between a stationary background and each new frame. Thus, this method relies on the feasibility to have a background at disposal. Nevertheless, in real cases this is often impossible. Usually, many algorithms depend on the feasibility of obtaining a background by using a short training sequence free of foreground objects. Nevertheless, in real cases this is often impossible. Few researches have been dedicated to the problem of background initialization (or *bootstrapping*, as we read in [1]) and none of them uses the approach we use.

The method we use allows generating a stationary background having at disposal a short sequence with an undetermined number of foreground objects. To be precise, our method aims to estimate for each pixel of every new frame of the model the intensity value to which that pixel has the maximum posterior probability. The earlier approach we used relied on the simple assumption that the maximum number of occurrences which a pixel value will get during the training period is due to the background. However, this approach required a long time in order to achieve a reliable background model, since it relies on a blind-update assumption. Actually, the method presented restricts the model update to non-moving pixels (selective-update). In this way, the new value of a stationary pixel, when such becomes "moving", does not alter the distribution for that pixel and usually the model strengthens its old background value. Although the method uses well known statistics, which are based on Bayes Theory, its application results in an original work.

This paper is organized as follows. Section 2 deals with the approaches actually used in most of the systems. The probabilistic framework is outlined in Section 3 and an accurate description of the initialization method we use follows in Section 4. Further on, extensive experiments accomplished on a challenging sequence are shown in Section 5 and they assess the effectiveness of the initialization model. At last, Section 6 draws conclusions.

2 Previous Works

Many probabilistic approaches in visual surveillance tasks require a background model before starting the sequence processing. Both the background generation and updating are crucial tasks. In order to cope with such problems, many authors have developed different methods, mostly are statistical. In this section we give an overview of the method used in most of the visual surveillance systems and of some interesting approaches not yet included in any system.

The work in [2] is a dated research but it represents one of the first attempts to reconcile statistic analysis and performance. In fact, here the background pixels are voted as the most frequent value during the image sequence. In [3] authors initialize the background by using the median intensity value for each pixel, thus relying on the assumption that each pixel will be visible for more than fifty percent of the time during the training sequence. In [4] authors implement a two-stage method which is able to obtain the background even in the presence of eventual moving foreground objects. In the first stage, a pixel-wise median filter over time is applied to several seconds of video in order to distinguish moving pixels from stationary pixels. In the second stage, only stationary pixels are updated. Further on, a pixel-based method updates the background model periodically and an object-based method updates the background to adapt to physical changes. The research in [5] presents an algorithm which is able to learn a model of the background when moving objects are present within the scene. Here, the input is a short monochromatic video sequence in which any number of moving objects may be present. Basically, authors use two methods, jointly. The first is called "adaptive smoothness method" ([6]) and starts by finding intervals of stable intensity. After that, a heuristic chooses the longest (i.e., the most stable) interval as the most suitable to represent the background. The second

*Address: Viale Risorgimento, 2, Bologna, ITALY 40136. E-mail: abevilacqua@deis.unibo.it

method uses the information about motion in proximity of the pixel: this will be discarded if any motion is toward itself. Finally, the statistical approach to background subtraction in [7] relies on an explicit model of illumination change and noise of pixel values. Authors formulate the detection problem of objects entering the scene as a statistical decision problem involving the relation between parameters of a reference image and the current frame. Here, they suppose that the two images are taken under different illumination conditions. In addition, the reference image has been taken when just background objects have been present.

3 Probabilistic Framework

Before entering the details of our method, let us recall the principles of the Bayes Theory applied to Computer Vision. In the following discussion, as in the remaining part of this work, an image I is represented by a 1-D vector which is nothing but the original 2-D image which has been sorted lexicographically. x_j is a pixel belonging to the image I, $N = |I|$ is the total number of pixels and i is a generic intensity gray level value.

Let $\mathbf{x}_j^T = [x_j^1 \cdots x_j^t \cdots x_j^T]$ be a vector of samples for the pixel x_j, $1 \leq j \leq N$, through frames $1, \cdots, T$, where T is the number of frames processed as far. Let $\mathbf{x}^t$ be the vector constituted by the first t samples, $1 \leq t \leq T$. The common form of Bayes Theory can be expressed by Eq. 1:

$$P(x_j^T|i) = \frac{p(i|x_j^T)P(x_j^T)}{p^T(i)} \tag{1}$$

By analyzing each probability concerning our problem by using histograms, we have:

$$p(i|x_j^T) = \frac{1}{T}h_{\mathbf{x}_j^T}(i) = \frac{1}{T}\sum_{t=1}^{T} h_{x_j^t}(i), \ \frac{1}{T}\sum_{i=0}^{G-1} h_{\mathbf{x}_j^T}(i) = 1 \tag{2}$$

$$P(x_j^t) = \frac{1}{N}, \qquad \forall t, \forall j \tag{3}$$

$$p^T(i) = \frac{1}{NT}\sum_{j=0}^{N-1} h_{\mathbf{x}_j^T}(i). \tag{4}$$

Here, $h_{\mathbf{x}_j^T}(i)$ is the temporal histogram of $\mathbf{x}_j^T$ related to the intensity gray level i. $p(i|x_j^T)$ is the *conditional probability density function (pdf)* for a given x_j at time T to have the intensity value i and it is called the *likelihood* of x_j with respect to i, at time T. This term indicates that, *prior probabilities* $P(x_j^T)$ of choosing the pixel x_j^T being equal, position x_j for which at time T $p(i|x_j^T)$ being the largest is the most "likely" to be the true position. At last, $p^T(i)$ is the prior pdf to have at time T an intensity value i. The left side of Eq.(1) is called *a posteriori (posterior) probability* and it shows that by observing the value i on the pixel x_j^T we can convert the prior probability $P(x_j^T)$ to this posterior probability.

4 Background initialization

In our algorithm, the input is taken by means of a stationary camera and it is a gray level sequence of few seconds, with an indefinite number of moving objects (e.g., cars and moving trees). The output of this bootstrapping sequence is an output model describing the static parts of the scene.

The basic idea which primarily inspired this method is that, during a reasonably long training sequence the most occurring value for the pixel x_j at time $T-1$ should reliably predict the most unchanged value at time T, i.e., a likely background value. Namely, we must find the intensity value i which the pixel x_j^{T-1} has the maximum likelihood to.

However, if one could know which pixels of each frame should be included in the background model B at time T, or better, which pixels *reliably should not be*, it would be enough to let them out in order to faster reflect the real distribution of the background values. Practically speaking, this means to estimate the posterior probability of x_j^T with respect to the intensity value i. The problem is described by expression (5):

$$B^T(x_j) = P(x_j^T|i) = \max_i(P(x_j^{T-1}|i)) \tag{5}$$

where $P(x_j^{T-1}|i)$ is the posterior probability of Eq.(1) and both the distributions $p(i|x_j^{T-1})$ and $p^{T-1}(i)$ are restricted to non-moving pixels. Strictly speaking, computation can be simplified since x_j has the same prior probability $P(x_j^t)$ for each t and j. When simplified, Eq.(1) becomes (using histograms and considering the first $T-1$ frames):

$$P(x_j^{T-1}|i) = \frac{h_{\mathbf{x}_j^{T-1}}(i)}{\sum_{j=0}^{N-1} h_{\mathbf{x}_j^{T-1}}(i)} \tag{6}$$

Expression (5) states that estimating the value of $B^T(x_j)$ (namely, the background pixel x_j for the frame T) means finding the intensity value i to which a given x_j^T has the maximum posterior probability.

Non-moving pixels are determined through a "background detection" method and to this purpose any motion detection technique is good. It is worth remarking that the moving detection method could be rough and not have been previously tuned. In fact it should only reveal most of the moving pixels with a low missed detection rate, even though a lot of false signals could be detected as well. For example, to this purpose we use temporal frame differencing. Figure 1 shows both the thresholded result of the two-frame difference (top) and the binary image attained after applying our morphological operations ([8]) (bottom). We see how morphological operations enlarge the blobs of the previous image. Removing both noise and false stationary pixels is not a hard task if one is not interested in achieving well-defined moving targets. In fact, even though many background pixels are erroneously detected as moving, thus we can attain a binary image where the entire "black" regions represent only *true* background pixels.

What happens in case we use posterior probabilities of Eq.(6) within a blind-update model or even a model wrongly updated by considering false stationary pixels? It would not yield interesting results, since $p^t(i)$ "corrects" the likelihood in an erroneous way ([9]). Figure 2 shows what happens after 10 frames. The error of the model increases rather than diminishing. Which is the reason? Because in case of uniform priors and under the same likelihood pdf, the posterior

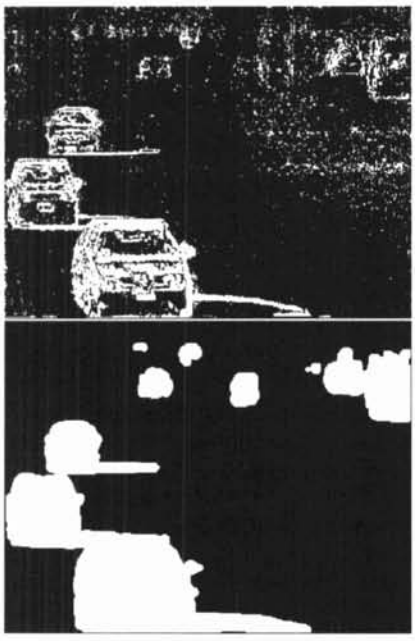

Figure 1: Motion detection achieved by means of the two-frame difference algorithm (top). The same image as above after having been roughly segmented through simple morphological operations (bottom).

probability rewards pixels having an intensity i with the lowest $p^t(i)$. We can explain this concept by a simple example in a better way (the order of magnitude of the used numbers does not refer to real cases). Let us suppose that the pixel $\mathbf{x}_j^t$ has the same likelihood $p(i|\mathbf{x}_j^t) = p(l|\mathbf{x}_j^t) = 0.025$ with respect to two different intensity values i, l belonging to the background and to the foreground distribution, respectively. Let i be representative for the background, i.e., it has accumulated a high number of occurrences over time, for example $p^t(i) = 0.06$. Let l belong to a foreground object and be representative for the foreground class. Since this distribution over time becomes wider with respect to the background distribution, usually $p^t(l) < p^t(i)$. Here, if we suppose $p^t(l) = 0.04$ then $P(\mathbf{x}_j^t|i) = 0.41$ and $P(\mathbf{x}_j^t|l) = 0.62$, hence $P(\mathbf{x}_j^t|l) > P(\mathbf{x}_j^t|i)$. To conclude, the more i is representative for the background the more l will have a larger posterior probability. In fact, in Figure 2, the foreground objects tend to persist within the model as the model becomes stronger. On the bottom right side of Figure 2, we still see the structure of the car which appeared in frame 2 (the first of the model, further shown in Figure 4, top). This discussion could not be true in case of a sequence where the static background is small with respect to larger foreground regions covering it with peaked distributions for most of the frames of the sequence. In any case, this situation could be not so usual in outdoor environments.

5 Experimental Results

The test sequence we use (a sample frame is shown in Figure 3, top) contains a cluttered daytime traffic sequence which has been sampled at 10 Hz and is of 210 frames. Images are 8-bit, gray level, with resolution of 384×288. The initialization algorithm has been written in ANSI C and works at 3 fps on a 800 MHz Pentium III PC.

Figure 2: Frame 10 of the model estimated with the maximum a posteriori probability without compensating for the different distribution of background and foreground pixels

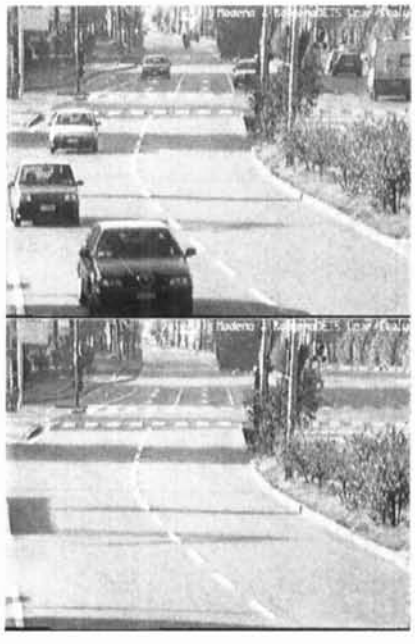

Figure 3: Frame 10 taken from the test sequence (top). Frame 106 generated by our background model

Figure 4 shows two distinct frames of the model we conceived. The black pixels point out moving foreground objects. The visual effect is that the background is revealing frame by frame as the model becomes more and more robust. The background image of Figure 4, bottom, shows a different behavior regarding vehicles which depends on their moving direction. While cars moving "towards the viewer" disappear faster, the background model still undergoes the effect of the vehicles which start from frame 2 (Figure 4, top) and move away on the right side of the highway. In fact, they are farther and free the background more slowly. Besides, by comparing frames 2 and 10 we see that the set of black pixels in frame 10 appears to be a subset of those of frame 2. Let us explain this behavior by defining two distinct situations. The first is when a pixel x_j is moving at time $t = 2$ and becomes station-

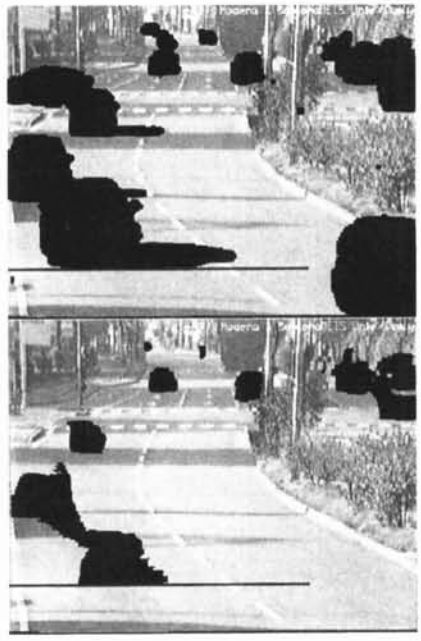

Figure 4: Frames 2 (top) and 10 (bottom) estimated with the computation of $p^t(i)$ and $p(i|\mathbf{x}_j^t)$ restricted to non-moving pixels. Black pixels in frame 2 appears to be a subset of those in frame 10

ary at time $t = k, k > 2$. Here we want to stress that if x_j is a moving pixel at time $t = 2$ (Figure 4, top) then $p(i|x_j^2) = 0, \forall i$. As soon as it becomes stationary, $p(i|x_j^k) > 0 \rightarrow p(\mathbf{x}_j^k|i) > 0$, hence the background is revealed. This is the reason why black areas shrink. The second circumstance occurs when a background pixel x_j is stationary at time $T-1$ and becomes subject to motion at time T. In this case, $p(i|\mathbf{x}_j^{T-1}) = p(i|\mathbf{x}_j^T)$; therefore changes in posterior probabilities are due only to $p^T(i)$. In fact, $p^T(i)$ differs from $p^{T-1}(i)$ in that $p^T(i)$ has accumulated occurrences of intensity values for non-moving pixels. Usually, $p^T(i)$ reinforces the model and $P(\mathbf{x}_j^T|i)$ slightly differs from its values in the previous frames, thus resulting in a reliable background value. This is the reason why black areas do not enlarge. We can verify the above analysis by considering the area below the blue line attached to the bottom side of the same car which persists in the foreground in frames 2 and 10 of Figure 4. Below this line, the background values are very similar, and this confirms previous considerations. To conclude, Figure 3, bottom, shows the background achieved by our model, after that a few more than one hundred frames have been processed. It is completely free of moving objects and it can be used as a reliable background in further background differencing algorithms.

6 Conclusions

An original background initialization model has been presented. It allows obtaining a background scene free of moving objects even in the presence of many moving targets. In addition, this model works also in the presence of a non-completely stationary background (e.g. showing waving tree phenomena).

Based on the Bayesian Theory, the method allows exploiting the "evidence" coming out of single frames before they are processed in order to speed up the method based on the likelihood probability. Practically speaking, information about true stationary pixels are employed in order to correct the likelihood probability and to use only reliable background pixels in order to build the model faster. A two-frame difference algorithm followed by simple morphological operations is enough to detect a reliable background. Therefore, simplicity joint to effectiveness makes such method apt to a wide number of scenes attained from different perspectives and illumination conditions.

Acknowledgements

We wish to thank Prof. Giorgio Baccarani and Prof. Riccardo Rovatti for their interest in the present work and for useful discussions. We also thank Prof. Luigi Di Stefano for having offered the sequence to study.

References

[1] K. Toyama, J. Krumm, B. Brumitt and B. Meyers, "Wallflower: Principles and Practice of Background Maintenance," Proceedings of the 7^{th} International Conference on Computer Vision, Corfu, Greece, (2) 255–261, 1999.

[2] A. Shio and J. Sklansky, "Segmentation of People in Motion," Proceedings of the IEEE Workshop on Visual Motion, 325–332, 1991.

[3] B. Gloyer, H. K. Aghajan, K. Y. Siu and T. Kailath, "Video-based freeway monitoring system using recursive vehicle tracking," Proceedings of IS&T-SPIE Symposium on Electronic Imaging: Image and Video Processing, 1995.

[4] I. Haritaoglu, D. Harwood and L. S. Davis, "A fast background scene modeling and maintenance for outdoor surveillance," Proceedings of the 15^{th} International Conference on Pattern Recognition, Barcelona, Catalonia, Spain, (4) 179–183, 2000.

[5] D. Gutchess, M. Trajković, E. Cohen-Solal, D. Lyons and A. K. Jain, "A Background Model Initialization Algorithm for Video Surveillance," Proceedings of the 8^{th} International Conference on Computer Vision, Vancouver, BC, Canada, (I) 733–739, 2001.

[6] W. Long and Y. H. Yang, "Stationary background generation: An alternative to the difference of two images," Pattern Recognition, 13(12) 1351–1359, 1990.

[7] D. Ohta, "A Statistical Approach to Background Subtraction for Surveillance Systems," Proceedings of the 8^{th} International Conference on Computer Vision, Vancouver, BC, Canada, (II) 481–486, 2001.

[8] A. Bevilacqua and M. Roffilli, "Robust desnoising and moving shadows detection in traffic scenes," Proceedings of the Technical Sketches of the IEEE Conference on Computer Vision and Pattern Recognition, Kauai Marriot, Hawaii, USA, 1–4, 2001.

[9] K.- P. Karmann and A. von Brandt, "Moving Object Recognition Using an Adaptive Background Memory," Time-Varying Image Processing and Moving Object Recognition, Elsevier Science Publisher B.V., 1990.