12—3

Adaptive background estimation

Mickael Pic, Luc Berthouze* and Takio Kurita
Neuroscience Research Institute
National Institute of Advanced Industrial Science and Technology, Japan

Abstract

Adaptive background techniques are useful for a wide spectrum of applications, ranging from security surveillance, traffic monitoring to medical and space imaging. With a properly estimated background, moving or new objects can be easily detected and tracked. Existing techniques are not suitable for real-world implementation, either because they are slow or because they do not perform well in the presence of frequent outliers or camera motion. We address the issue by computing a learning rate for each pixel, a function of a local confidence value that estimates whether a pixel is (or not) an outlier, and a global correlation value that detects camera motion. After discussing the role of each parameter, we report our experimental results, showing that our technique is fast but efficient, even in a real-world situation.

1 Introduction

The interest of separating dynamic objects, such as people, from a static background has been extensively discussed in the computer vision literature. Applications range from vehicle guidance [1], object tracking for security surveillance and traffic monitoring [2], environment description, image restoration [3], interactive games [4], to medical and space imaging and signal processing such as background estimation in experimental spectra [5]. Provided the background is static or slowly varying, segmentation algorithms can be divided in two types: interactive and automatic. Interactive techniques involve human interaction at least during the first frame. While they are flexible and accurate, they are not suitable for real-time and real-world applications for obvious reasons [6]. Automatic techniques are not straightforward because they cannot rely on a single source of information. Motion, for example, can be used to distinguish between background and foreground. However, in some applications such as videoconferencing and surveillance applications, objects may remain static for extended periods of time (e.g. a car stopped in a traffic jam). Color-based techniques are not very robust in applications where foreground and background share similar colors or when the environment is affected by changes in lighting conditions and noise in the camera. Thus, a combination of features is desirable [4].

To improve the quality of the background model, many techniques have been proposed that rely on the construction of a statistical background model, using mean value and standard deviation of Gaussian distributions - see [6, 7, 8] for a few examples. In all those techniques, a fixed adaptation rate is considered. Namely, for a pixel (or a disparity) value x_i, the value $\mu_i(t)$ of the i-th pixel of the estimated background will be given by $\mu_i(t) = \alpha x_i(t) + (1-\alpha)\mu_i(t-1)$.

In this paper, we suggest that the learning rate α should vary according to a confidence value at each pixel, in such a way that temporally present outliers be ignored, persistent outlier gradually become part of the background and significant background motion be rapidly learned. Such approach would yield increased flexibility in real-world applications. In traffic surveillance for example, a camera could be switched between different perspectives and rapidly adapt to the new background. Similarly, increased performance would be obtained in segmenting video streams or videoconferencing data with rapidly changing contexts.

2 Method

2.1 Confidences and correlations

The learning rate must take into account:

- The confidence value of a pixel with respect to its error with the estimated background (error above a statistical threshold);
- The overall correlation between a newly acquired frame and the estimated background. A trivial version would consist in computing the percentage of pixels whose error with the estimated background is above threshold.

The former has already been explored by one of the authors (T.K.) in previous work [9]. A pixel-wise difference $\epsilon_i^2(t)$ is computed between input image and background model as a quadratic error: $\epsilon_i^2(t) = (x_i(t) - \mu_i(t))^2$ where $x_i(t)$ is the value of pixel i at time t and $\mu_i(t)$ is the value of the estimated background pixel i at time t. A confidence value $\beta_i(t)$ is constructed with:

$$\beta_i(t) = exp\left(-\frac{\epsilon_i^2(t)}{2\sigma^2(t)}\right)$$

where $\sigma(t)$ is a robust estimation [10] of the standard deviation of the errors $\epsilon_i(t)$:

$$\sigma(t) = 1.4826\left(1 + \frac{5}{N-1}\right)\sqrt{d^2(t)}$$

where $d^2(t)$ denotes the median of all $\epsilon_i^2(t)$. The median is computed via an histogram of the errors. The

*Corresponding author. Address: Tsukuba AIST Central 2, Umezono 1-1-1, Tsukuba 305-8568 Japan. E-mail: Luc.Berthouze@aist.go.jp.

use of the median instead of the average proves valuable when only an object of small size is moved in the background. Note that, since we use a RGB color model of the background (as in [11] for example), all equations must be applied component-wise.

With respect to the determination of the correlation $\rho(t)$ however, we elected to work on the hue (H) component of the HSB model of both frames and estimated background so as to neglect the effects of illumination (variable brightness). $\rho(t)$ is then given by:

$$\rho(t) = \frac{\sum_i w_i(x_i(t) - \overline{x}(t))(\mu_i(t) - \overline{\mu}(t))}{\sqrt{\sum_i w_i(x_i(t) - \overline{x}(t))^2}\sqrt{\sum_i w_i(\mu_i(t) - \overline{\mu}(t))^2}}$$

where

$$\begin{aligned} \overline{x}(t) &= \sum_i w_i(t)x_i(t) \\ \overline{\mu}(t) &= \sum_i w_i(t)\mu_i(t) \\ w_i(t) &= \frac{\beta_i(t)}{\sum_i \beta_i(t)} \end{aligned}$$

Note that because $w_i(t)$ doesn't have any unit, its application to HSB components is reasonable. However, because the difference between hue components in the computation of ρ can lead to discontinuities (large difference between a hue component of 330 and 10), the following operator is used:

$$h_1 - h_2 \equiv \min(h_2 - h_1, h_1 - h_2 + 360) \text{ with } h_1 < h_2$$

2.2 Determination of the learning rate

For surveillance-type applications, for example, a robust background estimation must be obtained so that a robust segmentation of the moving objects can be achieved by subtracting the estimated background from the acquired image. Thus, pixels with low confidence value, e.g. because of temporal occlusion due to a moving person, should not necessarily be rapidly learned as being part of the background. Conversely, when the correlation is very low, e.g. rotation of the camera, learning must be rapid. Consequently, the learning rate $\alpha(t)$ must be some appropriate product of the two factors β and ρ. We define two functions f_ρ and f_β to weight the relative importance of the above factors in the evaluation of $\alpha(t)$.

For rapid adaptation to large changes of background (camera rotation for example), it is desirable that function f_ρ peaks when the correlation is low (with a threshold to be determined). In our experiments, f_ρ is given by:

$$f_\rho = \frac{atan(-k_\rho(\rho(t) - \rho_t))}{\pi} + \frac{1}{2} \qquad (1)$$

where ρ_t is the threshold separating large changes of background from local changes. The value of k_ρ determines the sensitivity of the function f_ρ to the threshold ρ_t. In our experiments, it was found that a very high value of k_ρ (from 500 to 1000) was reasonable for a threshold value ρ_t of 0.80. However, it was also found that depending on the speed or amplitude of the rotation, the minima in correlation ρ_t were spanned over too short a period of time, so that complete adaptation

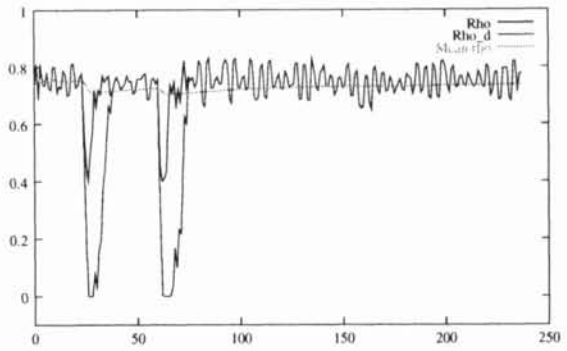

Figure 1: Illustration of the delay in the evaluation of the correlation: The threshold can be lower while the time of higher learning rate is extended.

was not possible. Consequently, a delay is introduced so that a high learning rate is maintained till the background is accurately learned. It is done by introducing a delay $\delta(t)$ and coupling it to a delayed correlation $\rho_d(t)$ defined as follows:

$$\begin{aligned} \delta(t+1) &= max(0, \delta(t) + q\,(\overline{\rho}(t) - \rho(t)) - p\rho(t)) \\ \rho_d(t) &= max(0, \rho(t) - \delta(t)) \end{aligned}$$

where q and p are accumulation (respectively decay) constants and $\overline{\rho}(t)$ is the running average of the correlation $\rho(t)$. The appropriate choice of the constants q and p enables a decrease of the threshold ρ_t (thus, improving the robustness of the system) and allows for complete adaptation following a camera rotation (see Figure 1). Naturally, $\rho_d(t)$ replaces $\rho(t)$ in Equation 1.

The local learning component of the learning rate is adjusted via function f_β so that the learning rate is higher when the confidence (that the pixel belongs to the background) is high. In the current state of our experiments, f_β is simply the identify function:

$$f_\beta = \beta_i(t)$$

A determination of $\alpha(t)$ as a simple product of $\alpha_c f_\beta(\beta_i(t)) f_\rho(\rho(t))$ is not satisfactory as it gives an equal weight to local changes and large scale motions. Instead, a strong weight to the correlation factor is desirable. It is obtained by using the following update law:

$$\alpha(t) = \alpha_{br} + \alpha_\rho f_\rho(\rho(t))\,(1 + f_\beta(\beta_i(t)))$$

where α_{br} is a learning base rate and α_ρ an weight factor determined so that the maximum learning rate of 1 is achieved when correlation $\rho(t)$ is minimal and $\beta_i(t)$ is maximal. In our experiments, the value of α_{br} (respectively α_ρ) was set to 0 (respectively 0.5).

3 Experimental results

We have tested the algorithm on various streams of images acquired by a camera monitoring automobile traffic on a major avenue. Such setup provides for an interesting case-study as (a) outliers (each passing car, motorbike or pedestrians) are numerous and moving at different speeds and (b) lighting conditions can vary quite significantly over extended periods of time. Figure 2 is a snapshot of the system while processing a

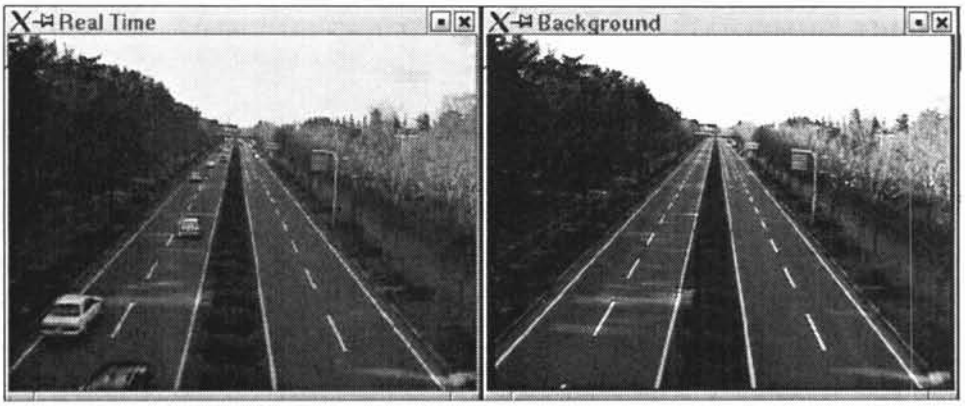

Figure 2: Snapshot of the system: (left) original frame, (right) estimated background.

frame. All computations are made on a single high-performance PC, in real-time. Provided that no initial outlier remains immobile over most of the experiment, the quality of the result obtained in this snapshot is independent of the initial conditions.

Figure 3 illustrates the effectiveness of our method to discriminate between outlier (low local confidence) and camera movement (low global correlation). The left-hand side column describes the behavior of the system in two cases: (a) pixel values vary either because of noise or because of the presence of outliers (e.g. $0 < t < 1000$); (b) an outlier appears and remains to the end of the experiment ($t > 1000$). In both cases, the learning rate remains almost unchanged, dampened by a very high correlation ρ. Thus, background adaptation subsequent to the introduction of a persistent outlier is very slow and (with this setting) is not completed before at least 500 frames. The right-hand side column describes the behavior of the system when the camera is panned (at time $t = 2020$). In this case, the correlation ρ decreases, which results in f_ρ increasing sharply, and remaining high over an extended period of time because of the accumulation of ρ_d. With the increase of f_ρ, α reaches a very high value and the background is re-learned in about 15 frames (at a frame rate of 30 fps). A side-effect of the accumulated ρ_d is that noise in subsequent estimations of the correlation is enhanced, which results in a learning rate remaining high for all pixels, even if locally the estimated background is correctly learned. Such problem can be solved with a proper selection of parameters k_ρ and ρ_t. Experiments carried out using various configurations of parameters show the system to be robust with a straightforward increase/decrease of the ratio between speed of learning of outlier and speed of adaptation to changes of context.

4 Conclusion

We reported a novel method for adaptive background estimation. Its originality resides in its simultaneous consideration of local confidence and overall correlation. Pixels with a low local confidence will be considered as outliers and as a result of which will see only a very low learning rate. Conversely, when the overall correlation is low, the system quickly adapts to compensate for either camera movements or rapid changes of context. The method was shown to be effective and robust in a real-world experimental setup.

References

[1] C. Ridder, O. Munkelt, and H. Kirchner. Adaptive background estimation and foreground detection using kalman-filtering. In *International Conference on Recent Advances in Mechatronics*, pages 193–199, 1995.

[2] D. Magee. Tracking multiple vehicles using foreground, background and motion models. In *ECCV Workshop on Statistical Methods in Video Processing*, pages 7–12, 2001.

[3] G. van Kempen and L. van Liet. Background estimation in nonlinear image restoration. In *Journal of the Optical Society of America A*, volume 17, number 3, pages 425–433, 2000.

[4] G. Gordon, T. Darell, M. Harville, and J. Woodfill. Background estimation and removal based on range and color. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–464, 1999.

[5] R. Fischer, K. Hanson, V. Dose, and W. von der Linden. Background estimation in experimental spectra. In *Physical Review E*, volume 61, number 2, pages 1152–1160, 2000.

[6] J. Pan, C. Lin, C. Gu, and M. Sun. A robust video object segmentation scheme with prerestored background information In *IEEE International Symposium on Circuits and Systems*, pages N/A, 2002.

[7] A. Francois and G. Medioni. Adaptive color background modeling for real-time segmentation of video streams. In *International Conference on Imaging Science, Systems, and Technology*, pages 227–232, 1999.

[8] C. Eveland, K. Konolige, and R. Bolles. Background modeling for segmentation of video-rate stereo sequences. In *IEEE Conference on Computer Vision and Pattern Recognition*, npages 266–272, 1998.

[9] T. Kurita, T. Takahashi, and Y. Ikeda. A neural network classifier for occluded images. In *International Conference on Pattern Recognition*, volume III, pages 45–48, 2002.

[10] P.J. Huber. *Robust Statistics*. John Wiley & Sons, 1981.

[11] S. Jabri, Z. Duric, H. Wechsler, and Z. Rosenfeld. Detection and location of people in video images using adaptive fusion of color and edge information. In *International Conference on Pattern Recognition*, pages 627–630, 2000.

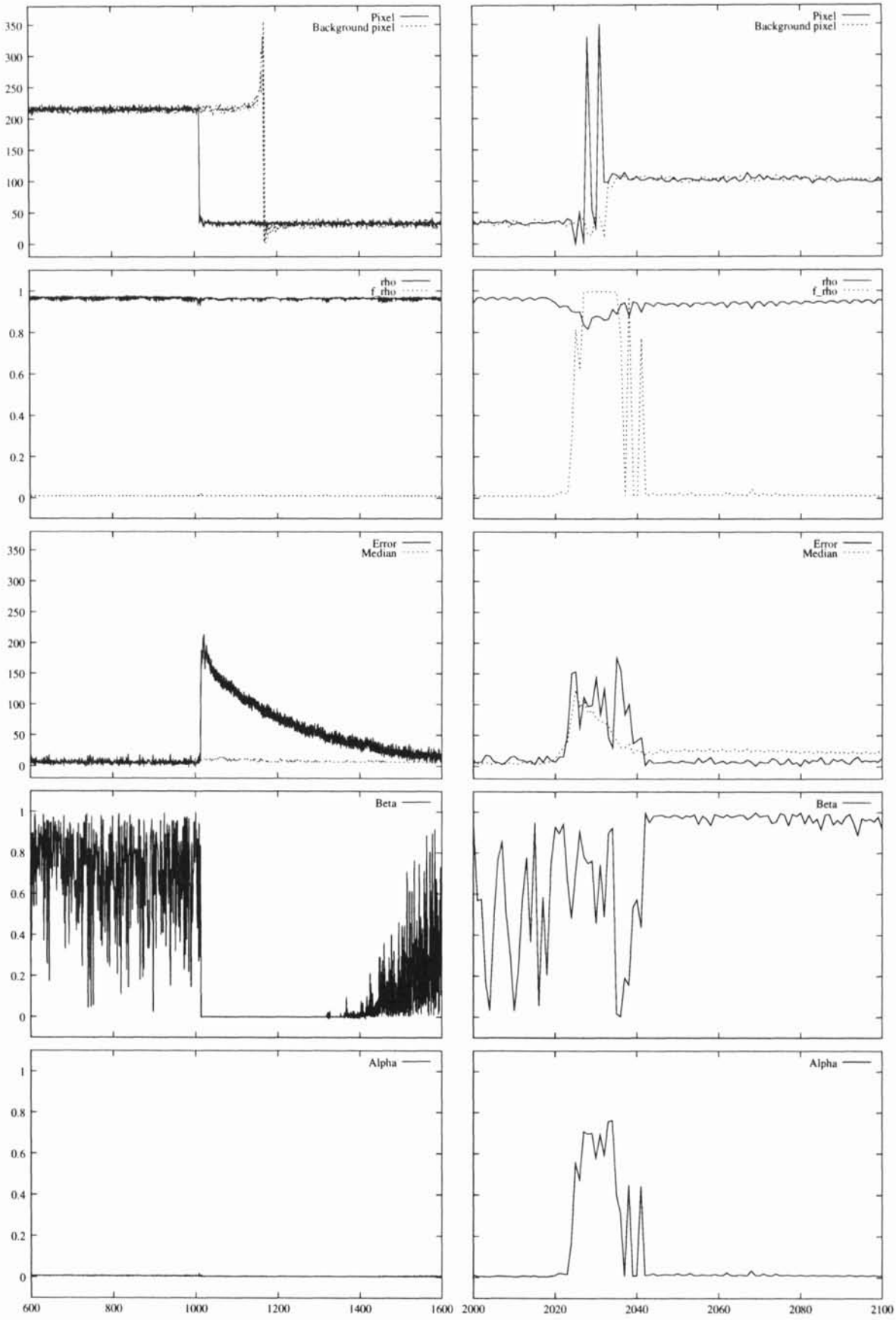

Figure 3: Discrimination of outlier (left column) versus panning camera (right column). From top to down, each frame denotes the time series of (a) a pixel value and the corresponding background pixel value, (b) the correlation ρ and f_ρ between consecutive frames, (c) the local error and the median of the error, (d) the confidence value β and (e) the computed learning rate α.