Adaptive background estimation: Computing a pixel-wise learning rate from local confidence and global correlation values

**Mickael PIC**, Luc BERTHOUZE, Nonmembers, and Takio KURITA, Regular Member

**SUMMARY** 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 experimental results, showing that our technique is fast but efficient, even in a real-world situation. Furthermore, we show that the same method applies equally well to a 3-camera stereoscopic system for depth perception.

**key words:** Background reconstruction, adaptive estimation, pixel confidence

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 [6]. Provided the background is static or slowly varying, segmentation algorithms can be divided into 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 [7]. 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, shadows and noise in the camera. While chromatic differences suitably address the shadow problem – shadows do not affect the chromatic component [10], it alone cannot overcome the effects of a changing illumination, which itself can be seen as a result of the interplay between surrounding lights and direct lights [5]. Thus, a combination of features is desirable [4].

To address some of these issues, techniques have been proposed that primarily rely on the construction of a statistical background model, using mean value and standard deviation of Gaussian distributions - see [7]-[9] 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

$$\mu_i(t) = \alpha x_i(t) + (1 - \alpha)\mu_i(t - 1)$$

(1)

In this paper, we suggest that the learning rate $\alpha$ 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

---

*Neuroscience Research Institute (AIST, Japan).
**Corresponding author.

---

This paper is an extended version of the paper presented at the IAPR Workshop on Machine Vision Applications, Nara (Japan), 2002 (pp. 451–454).
framework 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 in a previous study [12]. A pixel-wise difference $e_i^2(t)$ is computed between input image and background model as a quadratic error:

$$e_i^2(t) = (x_i(t) - \mu_i(t))^2 \tag{2}$$

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:

$$\beta_i(t) = \exp \left( -\frac{e_i^2(t)}{2\sigma^2(t)} \right) \tag{3}$$

where $\sigma(t)$ is a robust estimation [13] of the standard deviation of the errors $e_i(t)$:

$$\sigma(t) = 1.4826 \left( 1 + \frac{5}{N-1} \right) \sqrt{d^2(t)} \tag{4}$$

where $d^2(t)$ denotes the median of all $e_i^2(t)$. The median is computed via an histogram of the errors. The 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 [14] 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:

$$\rho(t) = \frac{\sum_i w_i(x_i(t) - \overline{x}(t))(\mu_i(t) - \overline{\mu}(t))}{\sqrt{\sum_i w_i(x_i(t) - \overline{x}(t))^2} \sqrt{\sum_i w_i(\mu_i(t) - \overline{\mu}(t))^2}} \tag{5}$$

where

$$\overline{x}(t) = \sum_i w_i(t)x_i(t) \tag{6}$$

$$\overline{\mu}(t) = \sum_i w_i(t)\mu_i(t) \tag{7}$$

$$w_i(t) = \frac{\beta_i(t)}{\sum_i \beta_i(t)} \tag{8}$$

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 $\rho$ can lead to discontinuities (large difference between a hue component of 330 and 10), the following operator is used:

$$h_1 - h_2 = \min(h_2 - h_1, h_1 - h_2 + 360) \text{ with } h_1 < h_2 \tag{9}$$

### 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 $\beta$ and $\rho$. We define two functions $f_\rho$ and $f_\beta$ 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 $f_\rho$ peaks when the correlation is low (with a threshold to be determined). In our experiments, $f_\rho$ is given by:

$$f_\rho = \frac{\tan(-k_\rho(\rho(t) - \rho_l))}{\pi} + \frac{1}{2} \tag{10}$$

where $\rho_l$ is the threshold separating large changes of background from local changes. The value of $k_\rho$ determines the sensitivity of the function $f_\rho$ to the threshold $\rho_l$. In our experiments, it was found that a very high value of $k_\rho$ (from 500 to 1000) was reasonable for a threshold value $\rho_l$ of 0.80. However, it was also found that depending on the speed or amplitude of the rotation, the minima in correlation $\rho_l$ were spanned over too short a period of time, so that complete adaptation 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 $\rho_{\delta}(t)$ defined as follows:

$$\delta(t+1) = \max(\delta(t) + q(\overline{\rho}(t) - \rho(t)) - pp(t)) \tag{11}$$

$$\rho_{\delta}(t) = \max(0, \rho(t) - \delta(t)) \tag{12}$$

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 $\rho_l$ (thus, improving the robustness of the system) and allows for complete adaptation following a camera rotation (see Figure 1). Naturally, $\rho_{\delta}(t)$ replaces $\rho(t)$ in Equation 10. The local learning component of the learning rate is adjusted via function $f_\beta$ 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, $f_\beta$ is simply the identify function:

$$f_\beta = \beta_i(t) \tag{13}$$

A determination of $\alpha(t)$ as a simple product of
\[ \alpha(t) = \alpha_{br} + \alpha_p f_p(\rho(t)) \left( 1 + f_\beta(\beta_i(t)) \right) \]  

where \( \alpha_{br} \) is a learning base rate and \( \alpha_p \) 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 \( \alpha_{br} \) (respectively \( \alpha_p \)) was set to 0 (respectively 0.5).

3. Experimental results

3.1 Traffic monitoring of a major avenue

We tested the algorithm on various streams of images acquired by a camera monitoring automobile traffic on a major avenue. Such setup provided for an interesting case-study as (a) outliers (each passing car, motorbike or pedestrians) were numerous and moving at different speeds and (b) lighting conditions could vary quite significantly over extended periods of time. Figure 2 is a snapshot of the system while processing a frame. All computations were made on a single high-performance PC, in real-time. Provided that no initial outlier remained immobile over most of the experiment, the quality of the result obtained in this snapshot was independent of the initial conditions. This issue of outliers occluding the background in the early stages of the estimation process has been dubbed initialization problem in [15]. The author reports a statistical method based on the improvement of a likelihood-based background model by using information about reliable stationary pixels through a simple motion detection algorithm. In future work, we could study whether such mechanism could be combined with our model to improve the early convergence of the estimation process.

Figure 4 illustrates the effectiveness of our method to discriminate between outlier (low local confidence) and camera movement (low global correlation as in Figure 3). 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
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 $\rho$ and $f_\rho$ between consecutive frames, (c) the local error and the median of the error, (d) the confidence value $\beta$ and (e) the computed learning rate $\alpha$. 

Fig. 4
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 \(\rho\). 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 \(\rho\) decreases, which results in \(f_\alpha\) increasing sharply, and remaining high over an extended period of time because of the accumulation of \(\rho_d\). With the increase of \(f_\rho\), \(\alpha\) 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 \(\rho_d\) 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 \(k_\rho\) and \(\rho_t\). 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.

3.2 Application to depth perception

In a second round of experiment, we applied the same algorithm to depth image sequences produced by a Triclops vision system (by Point Grey Research). This system uses three cameras in a “L” configuration, with stereo processing on both horizontal and texture information to extract a depth image from the environment. Because of the limited range of the device, we placed it in our office space and monitored the movements of co-workers and accessories being displaced (e.g., card-box, chairs). While lighting was not an issue (because of the predominance of direct light over surrounding light), limits in the sensors offered an interesting challenge. Indeed, black areas in the depth image could indicate either distant pixels, pixels for which depth could not be reliably computed and pixels located at the center of a large uniform area (no differential texture information). Since the confidence value computed in our method relies on pixel value, a noisy background estimation should be expected. Nevertheless, we only used a straightforward implementation of our method. Improvements of the method for such sensors will be published elsewhere.

The results were qualitatively similar to those obtained in color perception, in terms of speed of convergence and effectiveness of the model (see Figure 5 and Figure 6). Because the frames capture depth information, illumination and shadow effects were not a factor. However, areas of undetermined depth prevented a proper
A low local confidence will be considered as outliers and as a result of which will see only a very low estimation. Its originality resides in the simultaneous consideration of local confidence and overall correlation for robust background estimation.

We reported a novel method for adaptive background estimation. Its originality resides in the simultaneous consideration of local confidence and overall correlation for robust background estimation.

4. Conclusion

We reported a novel method for adaptive background estimation. Its originality resides in the simultaneous consideration of local confidence and overall correlation for robust background estimation.

Fig. 7 Treatment of outliers in the depth map. The top two frames correspond to the passage (at time step 850) of an outlier in the color map (left) and depth map (right). The middle frames correspond to the passage (at time step 1138) of another outlier. In the first case, the outlier is far enough from the triclops to allow for good depth measurements (white shadow in the left side of the frame). In the second case however, the outlier is too close to the triclops and the pixels corresponding to the outliers are undetermined (black shadow). The last two frames show the overall correlation $\rho$ (and $f_\rho$) over time. In both instances, the passage of the outlier triggers a change in the correlation (large changes) and the learning rate $\alpha$ at the pixel denoted by a white cross shows an increase (bottom right frame). Note that since the learning rate is magnified by a factor 300, the increase is in fact very small. It is also worth noting that during the passage of the second outlier, the confidence $\beta$ at the same pixel is not affected because the pixels are of undetermined values. However, because the overall correlation $\rho$ was below threshold, the learning rate increased nonetheless.

Identification of outliers, as shown by Figure 7. This came as a support to our claim (in introduction) that any single modality of information could not be sufficient for robust background estimation.

References


Mickael PIC received his DESS (equiv. MSc. degree) in Artificial Intelligence from the University Pierre&Marie Curie, Paris (France) in 2000. He joined the Neuroscience Research Institute, AIST of Japan, in 2001 as an assistant researcher. His research interests include, but are not limited to, computer vision, autonomous learning systems and new AI.

Luc BERTHOUGE received his Ph.D. in Computer Vision from the University of Evry (France) in 1996. He joined the Electrotechnical Laboratory (ETL) of Japan in 1995 as an invited researcher, and in 1996 as a permanent researcher. Since 2001, he is a research scientist at the Neuroscience Research Institute of AIST, Japan. He is also a Foreign Research Fellow at the University of Aizu, Japan. With research interests in developmental psychology, cognitive science, neuroscience and dynamical systems, he focuses on sensorimotor categorization and its role in the emergence of embodied cognition.

Takio KURITA received the B.Eng. degree from Nagoya Institute of Technology and the Dr. Eng. degree from the University of Tsukuba, in 1981 and in 1993, respectively. He joined the Electrotechnical Laboratory, AIST, MITI, Japan in 1981. From 1990 to 1991 he was a visiting research scientist at Institute for Information Technology, NRC, Ottawa, Canada. He is currently Deputy Director of Neuroscience Research Institute, National Institute of Advanced Industrial Science and Technology (AIST). His current research interests include statistical pattern recognition and neural networks. He is a member of IEICE, IEEE Computer Society, IPSJ, INNS, the Behaviormetric Society of Japan, and Japanese Academy of Facial Studies.