Object detection using spatio-temporal thresholding in image sequences

J.-H. Cho and S.-D. Kim

An algorithm using spatio-temporal thresholding for object detection with spatio-temporal distance metric in image sequences is proposed. The distance metric consists of the feature which uses the intensity and gradient at the same time in feature level instead of in decision level. In the model update process truncated variable adaptation rate is used, which can control adaptation rate up to its statistics, so it is able to maintain its statistics properly through the whole sequence. Some experimental results in various environments show that the averaged performance of the proposed algorithm is good.

Introduction: Object detection in image sequences has a very important role in many research areas which are related to computer vision. An object detector must be robust to some noise and be adaptive to illumination changes. In real-time applications it must run as fast as possible also. If the camera moves to track the object, generally global motion (GM) estimation and compensation must be carried out first. Previous algorithms have used temporal information [1]. In this Letter we focus on background subtraction, one of the most popular algorithms. There is a unified framework for background subtraction, which consists of the following three criteria. What kinds of features are used? What kind of distance metric is used to determine whether each pixel is object or background by thresholding? What is the adaptation rule? We propose some ideas with respect to each of the above criterion.

Statistical background model: We use a statistical background mosaic model having some advantages especially in the case of surveillance systems in which the camera moves with a regular pattern. To reduce accumulated errors, GM is estimated between the current image and the estimated background mosaic which contains no object, instead of the previous image [2]. After GM compensation, generally spatio-temporal statistics of the background model are assumed as in (1), and each parameter is defined as in (2):

$$\begin{aligned}&\mathbf{X}(x,y,t)\sim N(\mathbf{m}_t(x,y,t-1)\Sigma_t(x,y,t-1))\\&[\mathbf{X}(x,y,t)-\mathbf{m}_t(x,y,t-1)]\sim N(\mathbf{0},\Sigma_s(t))\end{aligned}\tag{1}$$

where $\mathbf{X}(x, y, t)$ is the extracted feature from the current image, which uses the intensity and gradient at the same time, and each subscript t and s means 'temporal' and 'spatial', respectively:

$$\begin{aligned}\mathbf{X}(x,y,t)&=[I(x,y,t)\ \mathbf{G}(x,y,t)^T]^T\\\mathbf{m}_t(x,y,t)&=[m_t^I(x,y,t)\ \mathbf{m}_t^{\mathrm{G}}(x,y,t)^T]^T\\\Sigma_t(x,y,t)&=\begin{bmatrix}\sigma_t^{I^2}(x,y,t)&\mathbf{0}^T\\\mathbf{0}&\Sigma_t^{\mathrm{G}}(x,y,t)\end{bmatrix}\\\Sigma_s(t)&=\begin{bmatrix}\sigma_s^{I^2}(t)&\mathbf{0}^T\\\mathbf{0}&\Sigma_s^{\mathrm{G}}(t)\end{bmatrix}\end{aligned}\tag{2}$$

where $\mathbf{G}(x, y, t)$ is gradient, which is estimated at (x, y, t) using the Sobel operator. The elements of $\Sigma_s(t)$ in (2) describe the spatial statistics of the background and are estimated as:

$$\begin{aligned}\sigma_s^{I^2}(t)&=\frac{1}{MN}\sum_{y=0}^{N-1}\sum_{x=0}^{M-1}[I(x,y,t)-m_t^I(x,y,t-1)]^2\\\Sigma_s^{\mathrm{G}}(t)&=\frac{1}{MN}\sum_{y=0}^{N-1}\sum_{x=0}^{M-1}[\mathbf{G}(x,y,t)-\mathbf{m}_t^{\mathrm{G}}(x,y,t-1)]\\&\quad[\mathbf{G}(x,y,t)-\mathbf{m}_t^{\mathrm{G}}(x,y,t-1)]^T\end{aligned}\tag{3}$$

where M and N are the width and height of the image, respectively.

Object detection using spatio-temporal thresholding: We make a decision whether each pixel is object or background by using the proposed spatio-temporal statistics. The spatio-temporal distance metric is estimated first after GM compensation as in (4), which is a kind of Mahalanobis distance. The decision is made by thresholding, i.e. if $Dist_t(x, y, t) > \gamma_t$ and $Dist_s(x, y, t) > \gamma_s$, the pixel is thought to be object, otherwise to be background. Each of these thresholds γ_t and γ_s is determined by the temporal and spatial statistics, respectively:

$$\begin{aligned}Dist_t(x,y,t)&=[\mathbf{X}(x,y,t)-\mathbf{m}_t(x,y,t-1)]^T\\&\quad\Sigma_t^{-1}(x,y,t-1)[\mathbf{X}(x,y,t)-\mathbf{m}_t(x,y,t-1)]\\Dist_t(x,y,t)&=[\mathbf{X}(x,y,t)-\mathbf{m}_t(x,y,t-1)]^T\\&\quad\Sigma_s^{-1}(t)[\mathbf{X}(x,y,t)-\mathbf{m}_t(x,y,t-1)]\end{aligned}\tag{4}$$

Adaptation with truncated variable adaption rate (TVAR): Background subtraction must be adaptive to noise and illumination changes to give an acceptable performance. In addition, we need to control adaption rate (AR) if necessary. Generally, we consider an adaptation rule as in (5):

$$\begin{aligned}\mathbf{m}_t(x',y',t)&=f(AR)\mathbf{m}_t(x',y',t-1)\\&\quad+(1-f(AR))\mathbf{X}(x',y',t)\\\Sigma_t(x',y',t)&=f(AR)\Sigma_t(x',y',t-1)+(1-f(AR))\\&\quad[\mathbf{X}(x',y',t)-\mathbf{m}_t(x',y',t)][\mathbf{X}(x',y',t)-\mathbf{m}_t(x',y',t)]^T\end{aligned}\tag{5}$$

where (x', y') is the background mosaic co-ordinate, so $\mathbf{X}(x', y', t)$ is the GM compensated feature and f is a function of AR, e.g. $f(AR) = e^{-AR}$. If $f(AR) = 1$, no adaptation occurs. If $f(AR) = 0$, there is an extremely fast adaptation. It is very important to find a suitable AR, but this is not always easy. A simple idea is that we choose AR, which is a function of $Dist_t(x', y', t)$, as in (6), i.e. we want to choose a high AR if $Dist_t(x', y', t)$ is very small, which means that the pixel is very likely to be background:

$$AR(x',y',t)=\frac{1}{Dist_t(x',y',t)}\tag{6}$$

where $Dist_t(x', y', t)$ is the GM compensated distance metric which is calculated at (x', y', t). The idea seems good, but because $Dist_t(x', y', t) \simeq 0$ in almost all background pixels, $AR(x, y, t)$ is too high to adapt the algorithm to illumination changes properly. Thus we propose the concept of TVAR as in (7) and Fig. 1:

$$TVAR(x',y',t)=\min\left\{\alpha,\max\left\{\frac{1}{\beta},\frac{1}{Dist_t}(x',y',t)\right\}\right\}\tag{7}$$

where α and β are truncation constants. Simulation results show that TVAR outperforms CAR or VAR.

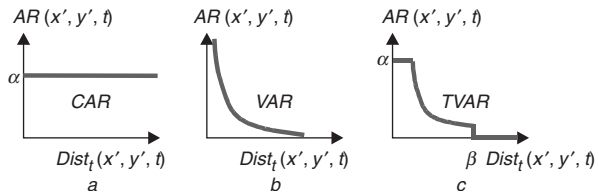


Fig. 1 *Adaptation rate (AR)*
a Constant AR (CAR)
b Variable AR (VAR)
c Truncated variable AR (TVAR)

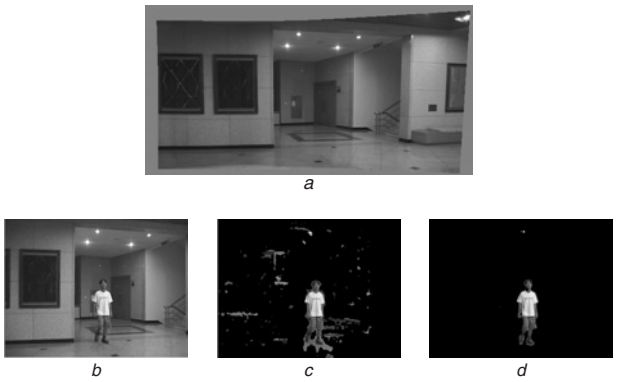
Experimental results: Some experimental results in various environments show that the averaged performance of the proposed algorithm is good. Fig. 2 shows the performance of the algorithms that use different features in a sequence. We know that feature $\mathbf{X}(x, y, t)$ outperforms the other two features. Even if the features $I(x, y, t)$ and $\mathbf{G}(x, y, t)$ are used at the same time in decision level, the performance is not as good as that of the proposed algorithm, which uses $\mathbf{X}(x, y, t)$ [3]. Fig. 3 shows the results which use spatio-temporal thresholding for object detection. The camera pans and tilts. Because the background mosaic is updated after each one detection cycle, false or misdetection errors may be accumulated and propagate through the whole sequence. The algorithm that uses the temporal statistics only cannot overcome this situation. Especially, there are many false detections in the region where the gradient is very strong, because we use $\mathbf{X}(x, y, t)$, which is considering gradient also. In Fig. 4 the results of object detection using each adaptation rule are shown in a whole image. Misdetection rate increases in the case of using CAR

because of ignoring its statistics. If we use VAR, false detection rate increases because of its extremely fast adaptation, i.e. when there is a small fluctuation in $\mathbf{X}(x, y, t)$, and is likely to be very large.

a *b* *c*

Fig. 2 *Feature level fusion of intensity and gradient*

a Object detection using $I(x, y, t)$
b Object detection using $\mathbf{G}(x, y, t)$
c Object detection using $\mathbf{X}(x, y, t)$



a *b* *c* *d*

Fig. 3 *Spatio-temporal thresholding for object detection*

a Background mosaic
b Current image
c Object detection using temporal statistics
d Object detection using spatio-temporal statistics

a *b* *c*

Fig. 4 *Object detection using TVAR*

a Object detection using CAR with $\alpha = 0.2$
b Object detection using VAR
c Object detection using TVAR with $\alpha = 0.2$, $\beta = 40$

20 May 2004
Electronics Letters online no: 20045316
doi: 10.1049/el:20045316

J.-H. Cho and S.-D. Kim (*Department of Electrical Engineering and Computer Science, Division of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), 373-1, Kusong-dong, Yusong-gu, Taejon 305-701, Republic of Korea*)

E-mail: mainkill@sdvision.kaist.ac.kr

References

1 Stauffer, C., and Grimson, W.E.L.: 'Adaptive background mixture models for real-time tracking'. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Fort Collins, CO, USA, Vol. 2, p. 252
2 Shum, H.-Y., and Szeliski, R.: 'Panoramic image mosaics', Tech. Rep. Microsoft Research, 1997 pp. 1–50
3 Gunatilaka, A.H., and Baertlein, B.A.: 'Feature-level and decision-level fusion of noncoincidently sampled sensors for land mine detection', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2001, **23**, (6), pp. 577–589