

Background Modelling in Demanding Situations with Confidence Measure

J. Rosell-Ortega G. Andreu-García A. Rodas-Jordà V. Atienza-Vanacloig
Grupo de Vision por Computador. DISCA. UPV. Spain
jarosell@doctor.upv.es, {gandreu, arodas, vatienza}@disca.upv.es

Abstract

Background subtraction is a popular technique in video surveillance. In order to use it, a background model must be built and updated according to light and scenario changes. We discuss in this paper a new algorithm (BAC) which creates or restores a background model based on the behaviour of pixels in successive frames, while performs a segmentation of objects in the scene yielding a confidence value for the obtained background, a problem which is addressed by few methods in the literature. This allows us to fulfil the requirement of producing a model, for instance in scenarios like airport halls, without interfering normal operation and still segment scenes.

1. Introduction ¹

Visual surveillance is an active research topic in computer vision and various surveillance systems have been proposed in recent years: [3], [4]. The visual surveillance process may be divided into the following steps: background modelling, motion detection, object classification, target tracking, behaviour understanding, human identification and data fusion.

Background subtraction is usually mentioned in the literature concerning visual surveillance as one of the most popular methods to detect regions of interest in frames. This technique consists in subtracting the acquired frame from a background model and classify as foreground all those pixels whose difference with the background is over a threshold. Thus, the importance of producing an accurate background model and choosing a precise threshold is obvious.

Background modelling involves generate a new background model at start and update it over time to cope with light changes. Several methods have been

proposed in the literature in order to create and maintain a background model, either by using statistical temporal functions on the most recent frames ([1], [5], [4]), or by using parametric or non-parametric mixture model of k Gaussians ([2]). The problem of model corruption is solved in some papers ([6]) by maintaining a database of background models and choose the most suitable for the situation. However this may take to maintain an extremely huge database with different models. None of the above mentioned methods are expected to perform any segmentation during their process.

Our application tracks people and luggage in an airport. The system consists of a set of cameras with a DSP running low-level vision algorithms covering the complete airport. We use background subtraction to detect targets, and it would be desirable being able to build background models without interfering the normal operation of the airport, even in high activity scenarios.

As a solution for this constraint, we propose an algorithm that allows object segmentation while the background model is built. Obviously, this segmentation will improve as background model's confidence increases.

Our goal is to be able to build a background model by taking frames no matter how many objects appear, permit a scene segmentation and avoid the need of storing background models. As a consequence of this step-by-step construction, it is possible to define a measure of the quality of the background model and the confidence of the foreground regions.

2 Background modelling with confidence

Background models are traditionally generated using statistical measures. In this paper, we propose not to use statistical properties of pixels, but their behaviour, to build the model.

Our algorithm considers consecutive grayscale frames $F(0), F(1), \dots, F(n)$, in which any pixel $p_{x,y} \in F(j)$ must belong either to foreground or to background. And builds a background model B starting

¹Acknowledgements: This work is supported partially by the sixth framework programme priority IST 2.5.3 Embedded systems. Project 033279.

from a frame $F(i), i \geq 0$. In this first frame it is impossible to classify pixels as background or foreground, as no further information is given. To decide which pixels may be used to update the background model and which not, a new similarity and motion criteria is defined in next sections.

Also, a confidence value is calculated for each pixel $b \in B$, in order to evaluate the security which this pixel is classified as belonging to background. The higher the confidence of the model, the better the background model.

A sample sequence may be seen in figure 1; images in columns show the model, frame and the result of the segmentation for different frames (1, 90 and 390).

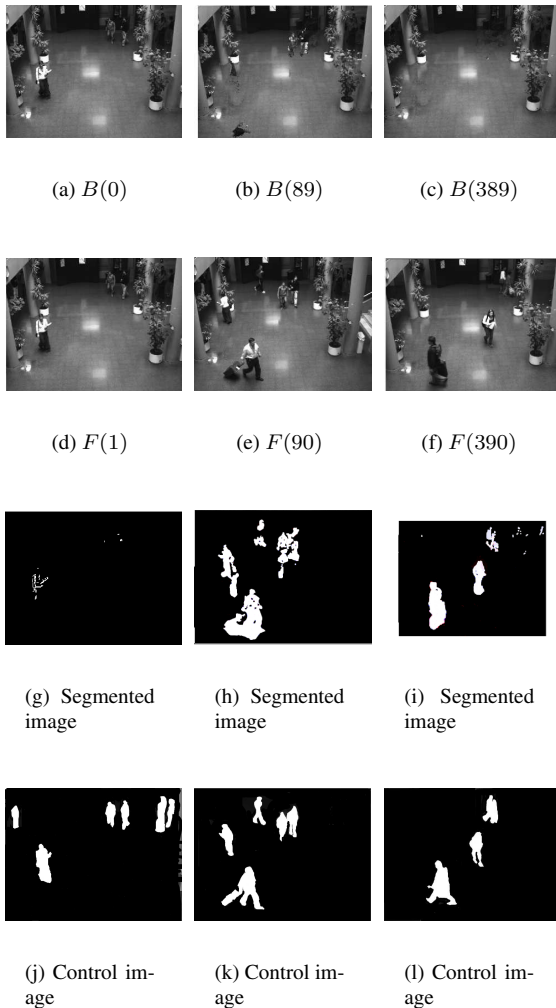


Figure 1: Sequence of background model reconstruction. From left to right, in columns, model, frame, automatic and hand segmentation for frames in $t = 1, 90$ and 390 .

2.1 Similarity

Similarity between two pixels is usually tested by comparing the difference of their gray levels with a threshold. We propose to translate into a function the intuitive idea behind "very similar" or "similar" by using a continuous function defined as $S(p, q) = e^{-\frac{|p-q|}{20}}$, $\mathfrak{R}^+ \rightarrow [0, 1]$ where p and q are gray levels of two pixels. This way, a difference degree and not an absolute value is calculated for pixels similarity.

2.2 Background algorithm with confidence

The background algorithm with confidence (BAC) starts by taking a frame $F(i)$ to be the initial background model $B(i)$ (the model in time i), and sets $\forall b_{x,y} \in B, c_{x,y}(i) = 0 \wedge \sigma_{x,y}(i) = 0$, being $c_{x,y}(i)$ the confidence value of pixel b and $\sigma_{x,y}(i)$ the filtered probability in time i .

Next two frames are ignored and used only to detect motion in the third frame. For all the next incoming frames $F(t)$, motion and similarities with $B(t-1)$ are sought for.

Motion in the scene is detected by considering similarity with previous frames. Being $q_{x,y} \in F(t)$ a pixel in the current frame, $p_{x,y} \in F(t-1)$ and $r_{x,y} \in F(t-2)$; we define the motion of $q_{x,y}$ as: $M(q) = \frac{(1-S(p_{x,y}, q_{x,y})) + (1-S(r_{x,y}, q_{x,y}))}{2}$. Similarity with the background is then given by $S(q_{x,y}, b_{x,y})$, being $b_{x,y} \in B(t)$.

We define then the probability that any pixel $q \in F(t)$ belongs to foreground as $P_{fore}(q) = \max(M(q), 1 - S(q, b))$ because pixels will belong to foreground if either their motion value is high or their difference with the background is high. On the other side, $P_{back}(q) = \min(1 - M(q), S(q, b))$ defines the probability that a pixel $q \in F(t)$ belongs to background if both its motion value is low and its similarity to current background is high (as stated in the constraints shown in section 2).

It must be noted that $P_{back} + P_{fore}$ not necessary should equal to 1. This is due to the fact that pixels may exist that are static for a short period, not long enough to be included into the background; and also pixels whose movement can not be appreciated by comparing two frames.

The model probabilities are then updated $\forall b \in B(t) : \sigma_b(t) = \frac{\sigma_b(t-1) \times c_b(t) + P_{back}(b)}{c_b(t) + 1}$. Being $\sigma_b(t)$ the filtered certainty of a pixel of belonging to background in time t . This avoids that an object captured in the first model stays for ever in the model.

Once $F(t)$ is segmented, we must update the model $B(t-1)$ to obtain $B(t)$, using pixels in $F(t)$. All pixels

Table 1: Percentage of pixels found for each hand-segmented target in control frames 90 and 390. Targets are not the same in both frames.

Target	Frame 90			Frame 390		
	BAC	mean	c_{target}	BAC	mean	c_{target}
1st target	0.69	0.74	0.946	0.96	0.88	0.9971
2nd target	0.90	0.70	0.977	0.89	0.71	0.996
3rd target	0.37	0.49	0.986	0.51	0.69	0.9971
4th target	0.47	0.49	0.933	-	-	-

$b_{x,y} \in B(t-1)$ are not updated in the same way, it depends on $P_{back}(b_{x,y})$, $c_{x,y}(t-1)$ and $\sigma_{x,y}(t)$.

Pixels are grouped in four different sets: pixels which belong to foreground ($fSet$), pixels which belong to background ($bSet$), pixels labelled as doubtful ($dSet$) and pixels in $B(t)$ whose gray level will be replaced by the gray level of pixels in $F(t)$ ($cSet$). They can be expressed as:

$$\begin{aligned}
 fSet &= \{p_{x,y} \in F(t) : P_{fore}(p_{x,y}) > 0.7\} \\
 dSet &= \{p_{x,y} \in fSet : \sigma_{x,y}(t) < 0.8 \wedge c_{x,y}(t-1) \geq 0.8\} \\
 cSet &= \{p_{x,y} \in fSet : \sigma_{x,y}(t) < 0.8 \wedge c_{x,y}(t-1) < 0.8\} \\
 bSet &= \{p_{x,y} \in F(t) : p_{x,y} \notin fSet\}
 \end{aligned}$$

Values defining the previous sets, were chosen to be very restrictive. The regions of interest of frame $F(t)$ are then defined by $fSet$. Let $\kappa = bSet \cup dSet$, $q_{x,y} \in F(t)$, the model $B(t)$ is updated as follows:

$$\begin{aligned}
 \forall b_{x,y} \in cSet : b_{x,y}(t) &= q_{x,y}(t) \\
 \forall b_{x,y} \in \kappa : b_{x,y}(t) &= q_{x,y}(t) + \alpha_{x,y}(t)(b_{x,y}(t-1) - q_{x,y}(t))
 \end{aligned}$$

The confidence of pixels in $B(t)$ is updated according to their classification:

$$\begin{aligned}
 \forall p_{x,y} \in bSet : c_{x,y}(t) &= c_p(t-1) + 1 \\
 \forall p_{x,y} \in dSet : c_{x,y}(t) &= c_p(t-1) - 1 \\
 \forall p_{x,y} \in cSet : c_{x,y}(t) &= 0
 \end{aligned}$$

The adaptation coefficient of background pixels is calculated as :

$$\begin{aligned}
 \forall p_{x,y} \in \kappa : \alpha_{x,y}(t) &= 0.98 \times \frac{c_{x,y}(t)}{c_{x,y}(t) + 1} \\
 \forall p_{x,y} \in cSet : \alpha_{x,y}(t) &= 0
 \end{aligned}$$

Together with its gray level value, each pixel $b_{x,y} \in B(t)$ provides a confidence value which may be used to ponder the quality of the segmentation. We define the segmentation confidence of the model $B(t)$ as $sc = \frac{1}{m*n} \times \sum \frac{c_b(t)}{c_b(t)+1}$, $\forall b \in B(t)$.

being $m \times n$ the number of pixels of the model. The segmentation confidence (sc) is calculated for a target T_i with a size l in pixels of frame $F(t)$, by particularizing this expression considering only the l pixels segmented for this target.

Finally, in order to test when the background model is stable the mean square quadratic difference ($msqd$) between two consecutive models is calculated; the algorithm finishes if the condition $msqd(B(t), B(t-1)) < 10^{-3} \wedge sc > 0.995$ verifies.

3 Experiments

We made experiments to test two different issues. First, several random frames from test videos were chosen as the base to reconstruct the background model. We then compared the background model obtained with the BAC algorithm, with the one obtained by using median, mean and mode with the same frames used by BAC. Next experiments were aimed to control how accurate the segmentation was, by using BAC to segment frames while the algorithm was under construction.

Videos from different sources were used with the aim of reproducing different situations; videos recorded by ourselves, real videos from the airport, Wallflower benchmark and AVSS benchmark. Videos had different lengths and were converted into grayscale when needed.

We compared the BAC's segmentation with a hand-made segmentation in order to obtain values of true positives (TP) are pixels classified as foreground which are foreground in the control imagen, true negatives (TN) are pixels classified as background in the image which are background in the control image. False positives (FP) and false negatives (FN) are defined as the complementary of the previous ones.

3.1 Results

Good results were obtained with BAC, they may be found together with resulting models using median, mean and mode applied to test videos at <http://www.vxc.upv.es/vision/proyectos/BAC>.

A representative situation aim of our developments is analysed in this section. The video starts in $F(0)$ with several people in a scene, simulating a surveillance system, in that moment $B(0)$ is created with targets with $sc = 0$, see figure 1(a). In order to evaluate quantitatively the evolution of BAC, we segmented manually 22 frames randomly selected.

In table 1, segmentation results for frames $F(90)$ and $F(390)$ obtained with BAC and mean are compared; sc of pixels found in each target's segmentation with BAC is shown under column " c_{target} ", for pixels not

Table 2: Results for Wallflower benchmark.

Video	TP	FP	TN	FN
camouflage	0.73	0.13	0.86	0.26
bootstrap	0.48	0.04	0.96	0.52
timeOfDay	0.36	0.01	0.99	0.64
wavingTree	0.73	0.26	0.74	0.27
movedObject	1	0.01	0.99	0
lightSwitch	0.44	0.02	0.98	0.56
foregroundAperture	0.51	0.05	0.95	0.49

correctly segmented, sc was under 0.001. Both frames may be also found in figure 1.

In $F(90)$, the four objects present in the scene are segmented with BAC and mean; only those dark objects which are far in the field of view of the camera are segmented more poorly (target 3); something similar happens with target 4, which is a group of two people moving still in the same area they occupied at the beginning of the movie. We consider that with at least 45% of the total size of pixels detected of a target is sufficient to continue with classification and tracking tasks, if they are grouped in an only blob.

In figure 1(a) and 1(b) $B(0)$ and $B(89)$ are shown, it may be seen that in $B(89)$, background model has achieved $c = 0.982$ and some targets have been removed. Improvement over time is evident as $B(389)$ contains no target. This improvement manifests in $F(390)$ with a better segmentation and a model with $c = 0.997$.

Evolution of BAC's confidence, TP and TN , of BAC and mean are shown in figure 2. Objects standing still for long periods of time influence negatively the value of TP . The plot shows that BAC segments correctly more pixels than mean. In $F(201)$ several objects leave the scene and others start coming in and in $F(680)$ some objects stand still; this explains some foreground pixels not found. On the other side, TN , easily reach a high level as area of quiet targets is small compared to the image.

Table 2 shows results for Wallflower benchmark. Values of TN are high, though videos were too short for BAC to converge. For sequence "wavingTree" sequence, fails due to the movement of the tree. In "lightSwitch", BAC was started in the moment lights were switched on. Finally, in "bootstrap", brightness makes BAC fail to find correctly the targets, though it find most of the pixels associated to them.

4 Conclusions

We introduced a different approach to background modelling. Our algorithm is aimed to reconstruct background models and permit segmentation and tracking

of objects. We tested the algorithm in several situations with test videos from different sources. We found it was difficult to analyse quantitatively the results and thus, we set a web-site where videos showing algorithm evolution is illustrated.

Experiments show that BAC obtains background models equal to those obtained by using any statistical technique; with the added benefit of permitting segmentation from the very beginning of the process; as was the goal and together with a confidence measure of the obtained model. Also, the experiments performed with the Wallflower test set result promising. Our efforts should be address in near future to improve the response of the algorithm to shadows and brightness. And extending this schema to other colour coordinates.

References

- [1] R. Cucchira, C. Grana, M. Piccardi, and A. Prati. Detecting objects, shadows and ghosts in video stream by exploiting colour and motion information. *ICIAP 2001*, pages 360 – 365, 2001.
- [2] A. Elgammal, D. Harwood, and L. Davis. Non-parametric model for background subtraction. *ECCV 2000*, pages 751 – 767, 2000.
- [3] I. Haritaoglu, D. Harwood, and L. S. Davis. W4: Real-time surveillance of people and their activities. *IEEE Transactions on PAMI*, pages 809 – 830, 2000.
- [4] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, man, and Cybernetics*, 2004.
- [5] B. Shoushtarian and H. Bez. A practical adaptive approach for dynamic background subtraction using an invariant colour model and object tracking. *Pattern Recognition Letters 26*, pages 5 – 26, 2005.
- [6] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. *7th Intl. Conf. on Computer Vision, Kerkyra, Greece*, pages 255–261, 1999.

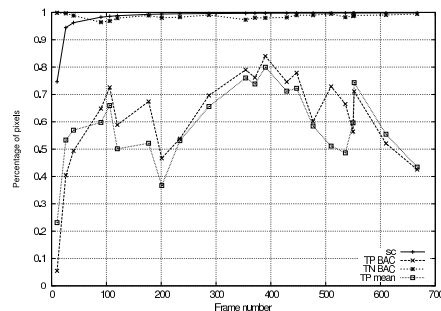


Figure 2: Evolution of confidence, TP and TN for the discussed video. Spots correspond to control frames.