Semantic Background Subtraction

Marc Braham, Sébastien Piérard and Marc Van Droogenbroeck
{M.Braham, Sebastien.Pierard, M.VanDroogenbroeck}@ulg.ac.be
University of Liège, Department of Electrical Engineering and Computer Science (Montefiore Institute), Belgium

Abstract

We introduce the notion of semantic background subtraction, a novel framework for motion detection in video sequences. The key innovation consists to leverage object-level semantics to address the variety of challenging scenarios for background subtraction. Our framework combines the information of a deep semantic segmentation network, expressed by a probability for each pixel, with the output of any background subtraction algorithm to reduce false positive detections produced by illumination changes, dynamic backgrounds, strong shadows, and ghosts. In addition, it maintains a fully semantic background model to improve the detection of camouflaged foreground objects. Experiments led on the CDNet dataset show that we managed to improve, significantly, almost all background subtraction algorithms of the CDNet leaderboard, and reduce the mean overall error rate of all the 34 algorithms (resp. of the best 5 algorithms) by roughly 50% (resp. 20%).

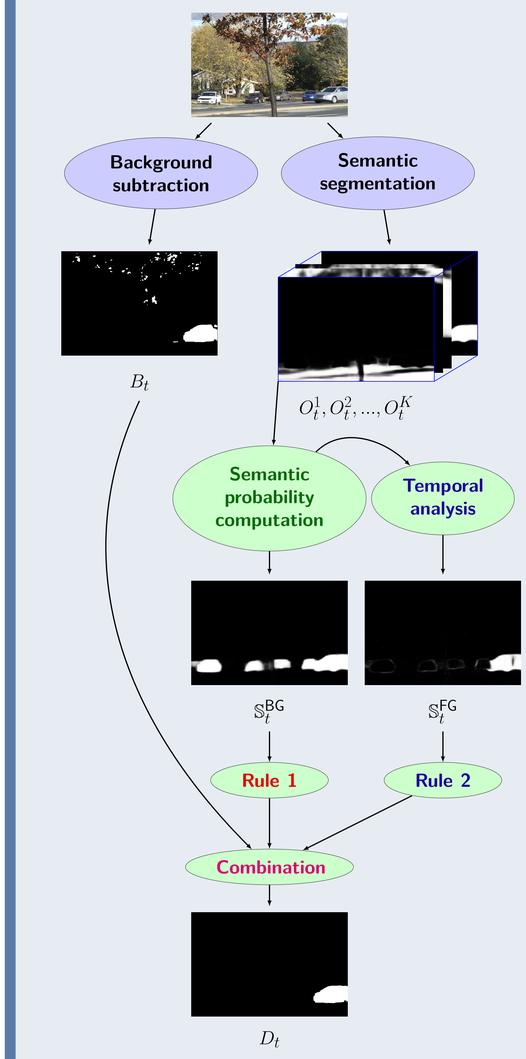

Figure 1 : We present a framework that improves the binary segmentation maps produced by background subtraction algorithms by leveraging object-level semantics provided by a semantic segmentation algorithm.

Motivation

Our objective is to show the possibility of leveraging state of the art semantic segmentation algorithms to improve the performance of BackGround Subtraction (BGS) algorithms, without modifying them or accessing their internal elements (e.g. their model and parameters). Our framework compensates for the errors of any BGS algorithm by combining, at the pixel level, its result $B \in \{\mathsf{BG}, \mathsf{FG}\}$ with two signals ($\mathbb{S}^{\mathsf{BG}}$ and $\mathbb{S}^{\mathsf{FG}}$) derived from the semantics, as shown in Figure 1. While the first signal supplies the information necessary to detect many BG pixels with high confidence, the second helps to detect FG pixels reliably. The result of the combination is denoted by $D \in \{\mathsf{BG}, \mathsf{FG}\}$.

Semantic probability computation

Let $C = \{c_1, c_2, ..., c_N\}$ be a set of N disjoint object classes. We assume that the semantic segmentation algorithm outputs a real-valued vector $\boldsymbol{v}_t(x) = [v_t^1(x), v_t^2(x), ..., v_t^N(x)]$, where $v_t^i(x)$ denotes a score for class c_i at the pixel location x at time t. The probabilities $p_t(x \in c_i)$ are estimated by applying a $\mathrm{softmax}$ function to $\boldsymbol{v}_t(x)$. Our framework requires the definition of the subset R ($R \subset C$) of all object classes semantically relevant for motion detection problems. The semantic probability, which is the probability for a particular pixel to belong to an object of interest, is defined and computed as

$$p_{S,t}(x) = p_t(x \in R) = \sum_{c_i \in R} p_t(x \in c_i)$$

Leveraging semantics to detect background pixels

.

It is possible to leverage semantics to detect background, as all pixels with a low semantic probability value $p_{S,t}(x)$ should be labeled as background, regardless of the decision $B_t(x)$. Therefore, we compare the signal $\mathbb{S}_t^{\mathsf{BG}}(x) = p_{S,t}(x)$ to a decision threshold τ_{BG}, as given by rule 1:

$$\textbf{Rule 1}: \quad \mathbb{S}_t^{\mathsf{BG}}(x) \leq \tau_{\mathsf{BG}} \rightarrow D_t(x) = \mathsf{BG} \ . \tag{1}$$

Rule 1 provides a simple way to address the challenges of illumination changes, dynamic backgrounds, ghosts, and strong shadows, which severely affect the performances of BGS algorithms by producing many false positive detections.

Leveraging semantics to detect foreground pixels

In order to help detecting the foreground, we have to use $p_{S,t}(x)$ in a different way than for rule 1, as semantically relevant objects may be present in the background (*e.g.* a car parked since the first frame of the video). To account for this possibility, our solution consists to maintain a purely semantic background model for each pixel. More precisely, we denote by $M_t(x)$ the probability modeling the semantics of the background at the pixel x at time t. This model allows to detect large increases of $p_{S,t}(x)$, observed when a foreground object appears in front of a semantically irrelevant background (*e.g.* a car moving on a road or a pedestrian walking in front of a building). This leads us to the following decision rule:

$$\textbf{Rule 2}: \quad \mathbb{S}_t^{\mathsf{FG}}(x) \geq \tau_{\mathsf{FG}} \rightarrow D_t(x) = \mathsf{FG} \ , \tag{2}$$

with the signal $\mathbb{S}_t^{\mathsf{FG}}(x) = p_{S,t}(x) - M_t(x)$, and τ_{FG} denoting a second threshold. Rule 2 aims at reducing the number of false negative detections due to camouflage, *i.e.* when background and foreground share similar colors.

The BGS is used when semantics is not decisive

The semantic probability $p_{S,t}(x)$ alone does not suffice for motion detection. If conditions of rules 1 and 2 are not met, which means that semantics alone does not provide enough information to take a decision, we delegate the final decision to the BGS algorithm: $D_t(x) = B_t(x)$. The complete classification process is summarized in Table 1.

Complete classification process

$B_t(x)$	$\mathbb{S}_t^{\mathsf{BG}}(x) \leq \tau_{\mathsf{BG}}$	$\mathbb{S}_t^{\mathsf{FG}}(x) \geq \tau_{\mathsf{FG}}$	$D_t(x)$
BG	false	false	BG
BG	false	true	FG
BG	true	false	BG
BG	true	true	X
FG	false	false	FG
FG	false	true	FG
FG	true	false	BG
FG	true	true	X

Table 1 : Our combination of three signals for semantic BGS. Rows corresponding to "don't-care" values (X) cannot be encountered, assuming that $\tau_{\mathsf{BG}} < \tau_{\mathsf{FG}}$.

The importance of both rules should be emphasized. Rule 1 always leads to the prediction of BG, so its use can only decrease the *True Positive Rate* TPR and the *False Positive Rate* FPR, in comparison to the BGS algorithm used alone. To the contrary, rule 2 always leads to the prediction of FG, and therefore its use can only increase the TPR and the FPR. The objective of improving both the TPR and the FPR can thus only be reached by the joint use of both rules.

Experimental results

We applied our framework to the 34 BGS methods whose segmentation maps (which directly provide the binary decisions $B_t(x)$) are available on the website of the CDNet dataset for 53 video sequences organized in 11 categories. We rely on the recent deep architecture PSPNet trained on the ADE20K dataset to extract semantics. In order to show the effectiveness of our framework, we compare the performances of BGS methods applied with or without semantics. The improvement is defined as

$$\text{improvement} = \frac{\mathsf{ER}_{BGS} - \mathsf{ER}_{BGS+SEM}}{\mathsf{ER}_{BGS}} \ , \tag{3}$$

where ER denotes the mean *Error Rate* over a particular set of BGS methods and a set of categories from the CDNet dataset. Per-category improvements are detailed in Figure 2.

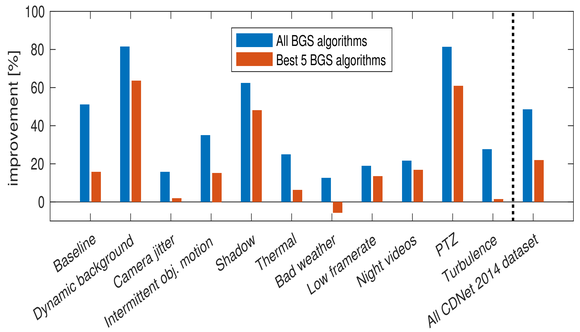

Figure 2 : Mean improvements (see (3)) of our framework (with default thresholds).

As illustrated in Figure 2, we observe huge improvements for "Baseline", "Dynamic background", "Shadow", and "PTZ" categories. We manage to reduce the mean overall error rate of all the 34 algorithms (resp. of the best 5 algorithms) by roughly 50% (resp. 20%). Figure 3 shows that our framework tends to reduce significantly the FPR of BGS algorithms, while increasing simultaneously their TPR.

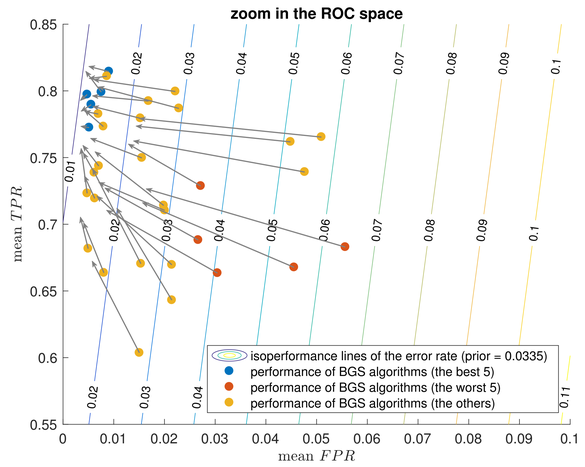

Figure 3 : Effect of our framework on the position of BGS classifiers in the overall ROC space of the CDNet dataset, with default thresholds. It tends to reduce the FPR significantly, while simultaneously increasing the TPR.

Figure 4 illustrates the benefits of our semantic background subtraction framework for several challenging scenarios of real-world video sequences. It reduces drastically the number of false positive detections caused by dynamic backgrounds, ghosts, and strong shadows, while mitigating simultaneously color camouflage effects.

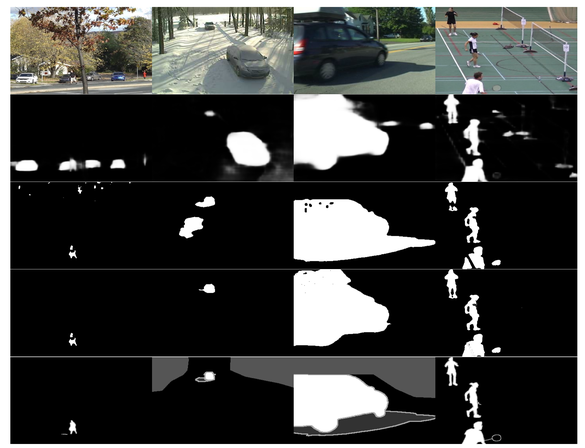

Figure 4 : Our framework addresses robustly dynamic backgrounds (column 1), ghosts (column 2), strong shadows (column 3) and camouflage effects (column 4). From top row to bottom row: the input image, the probabilities $p_{S,t}(x)$, the output of IUTIS-5, the output of IUTIS-5 integrated in our framework, and the ground truth.

Conclusion

We have presented a novel framework for motion detection in videos that combines BackGround Subtraction (BGS) algorithms with two signals derived from object-level semantics extracted by semantic segmentation. The framework is simple and universal, *i.e.* applicable to every BGS algorithm, because it only requires binary segmentation maps. Experiments led on the CDNet dataset show that we managed to improve significantly the performance of 34 BGS algorithms, by reducing their mean overall error rate by roughly 50%.