SEMANTIC BACKGROUND SUBTRACTION

M. Braham, S. Piérard and M. Van Droogenbroeck

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

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 semantic segmentation algorithm, 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%). Note that a C++ implementation of the framework is available at http://www.telecom.ulg.ac.be/semantic.

***Index Terms*—** background subtraction, change detection, semantic segmentation, scene labeling, classification

1. INTRODUCTION

Background subtraction is a popular approach for detecting moving objects in video sequences. The basic idea consists in comparing each video frame with an adaptive background model (which can be reduced to a single image) free of moving objects. Pixels with a noticeable difference are assumed to belong to moving objects (they constitute the *foreground*) while others are classified as *background*.

Over the last two decades, a large number of methods have been proposed for this task (see [1, 2] for reviews). Most of them model the background using low-level features such as color components [3, 4], edges [5], texture descriptors [6], optical flow [7], or depth [8]. A comprehensive review and classification of features used for background modeling can be found in [9]. While most of these features can be computed with a very low computational load, they cannot address simultaneously the numerous challenges arising in real-world video sequences such as illumination changes, camouflage, camera jitter, dynamic backgrounds, shadows, etc. Upper bounds on the performance of pixel-based methods based exclusively on RGB color components were simulated in [10]. In particular, it was shown that these algorithms fail to provide a perfect segmentation in the presence of noise and shadows, even when a perfect background image is available.

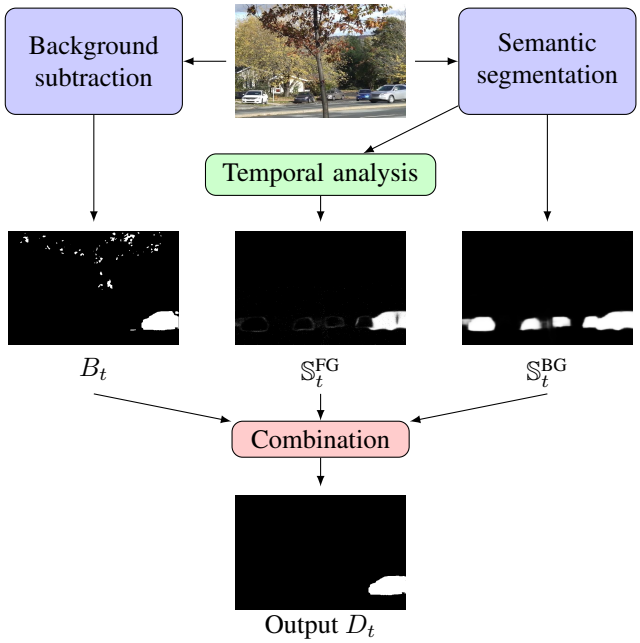

Fig. 1. We present a framework that improves the binary segmentation maps produced by background subtraction algorithms by leveraging object-level semantics provided by a semantic segmentation algorithm (see Section 2).

Our solution consists in the introduction of semantics. Humans can easily delineate relevant moving objects with a high precision because they incorporate knowledge from the semantic level: they know what a car is, recognize shadows, distinguish between object motion and camera motion, etc. The purpose of *semantic segmentation* (also known as *scene labeling* or *scene parsing*) is to provide such information by labeling each pixel of an image with the class of its enclosing object or region. The task is difficult and requires the

simultaneous detection, localization, and segmentation of semantic objects and regions. However, the advent of deep neural networks within the computer vision community and the access to large labeled training datasets have dramatically improved the performance of semantic segmentation algorithms [11, 12, 13, 14]. These improvements have motivated their use for specific computer vision tasks, such as optical flow estimation [15]. In this paper, we leverage object-level semantics for motion detection in video sequences and present a generic framework to improve background subtraction algorithms with semantics.

The outline of the paper is as follows. We describe the details of our semantic background subtraction framework in Section 2. In Section 3, we apply our proposed approach to all the 34 background subtraction methods whose segmentation maps are available on the website of the CDNet dataset [16] (named CDNet hereafter) and discuss the results. Finally, Section 4 concludes the paper.

2. SEMANTIC BACKGROUND SUBTRACTION

Our framework compensates for the errors of any background subtraction (named BGS hereafter) algorithm by combining, at the pixel level, its result $B \in \{\text{BG}, \text{FG}\}$ with two signals ($\mathbb{S}^{\text{BG}}$ and $\mathbb{S}^{\text{FG}}$) 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 $D \in \{\text{BG}, \text{FG}\}$. Our objective is to show the possibility of leveraging state of the art semantic segmentation algorithms to improve the performance of most BGS algorithms, without modifying them or accessing their internal elements (*e.g.* their model and parameters).

2.1. Leveraging semantics to detect background pixels

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 $\boldsymbol{v}_t(x)$. Let R ($R \subset C$) be the subset of all object classes semantically relevant for motion detection problems. The semantic probability is defined and computed as $p_{S,t}(x) = p_t(x \in R) = \sum_{c_i \in R} p_t\,(x \in c_i)$.

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 $\mathbb{S}_t^{\text{BG}}(x) = p_{S,t}(x)$ to a decision threshold τ_{BG}, as given by rule 1:

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

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. The optimal value of τ_{BG} is related to the performance of the BGS algorithm for the BG class, as explained in Section 3.

2.2. 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. Typical initialization and updating steps of this semantic model can be the following:

$$\begin{cases} M_0(x) = p_{S,0}(x) \\ D_t(x) = \text{FG} \rightarrow M_{t+1}(x) = M_t(x) \\ D_t(x) = \text{BG} \rightarrow_{\alpha} M_{t+1}(x) = p_{S,t}(x) \\ \qquad\qquad\quad \rightarrow_{1-\alpha} M_{t+1}(x) = M_t(x) \end{cases} \tag{2}$$

with $\rightarrow_{\alpha}$ denoting a probability α of application (α is arbitrarily set to 0.00024 in our experiments). This conservative updating strategy was introduced in [4] to avoid model corruptions due to intermittent and slow moving objects. The semantic background 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:

$$\text{rule 2:}\quad \mathbb{S}_t^{\text{FG}}(x) \geq \tau_{\text{FG}} \rightarrow D_t(x) = \text{FG}\ , \tag{3}$$

with the signal $\mathbb{S}_t^{\text{FG}}(x) = p_{S,t}(x) - M_t(x)$, and τ_{FG} denoting a second threshold, whose optimal value is related to the performance of the BGS algorithm for the FG class, as explained in Section 3. Rule 2 aims at reducing the number of false negative detections due to camouflage, *i.e.* when background and foreground share similar colors.

2.3. The BGS is used when semantics is not decisive

The semantic probability $p_{S,t}(x)$ alone does not suffice for motion detection. This is illustrated by the case in which a semantically relevant object (*e.g.* a car in the foreground) moves in front of a stationary object of the same semantic class (*e.g.* a car parked in the background). The semantic probability $p_{S,t}(x)$ being the same for both objects, it is impossible to distinguish between both. 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.

$B_t(x)$	$\mathbb{S}_t^{\text{BG}}(x) \leq \tau_{\text{BG}}$	$\mathbb{S}_t^{\text{FG}}(x) \geq \tau_{\text{FG}}$	$D_t(x)$
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 $\tau_{\text{BG}} < \tau_{\text{FG}}$.

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.

3. EXPERIMENTAL RESULTS

We applied our framework to all 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 [16] for 53 video sequences organized in 11 categories. We rely on a recent deep architecture, PSPNet [13] (ranked 1st in the PASCAL VOC 2012 object segmentation leaderboard [17] on the 6th of February 2017), trained on the ADE20K dataset [18] to extract semantics, using a publicly available model [19]. The last layer of the model provides a real value in each pixel for each of 150 object classes of the ADE20K dataset (C). Our subset of semantically relevant objects is R = {person, car, cushion, box, book, boat, bus, truck, bottle, van, bag, bicycle}, corresponding to the semantics of CDNet foreground objects.

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

$$\text{improvement} = \frac{\text{ER}_{BGS} - \text{ER}_{BGS+SEM}}{\text{ER}_{BGS}} \quad , \qquad (4)$$

where ER denotes the mean *Error Rate* over a particular set of BGS methods and a set of categories from the CDNet dataset. We considered three policies to set τ_{BG} and τ_{FG}.

(1) Optimization based policy. First, we performed a grid search optimization, for each BGS algorithm specifically, to select the thresholds producing the best overall F score:

$$\left(\tau_{\text{BG}}^{opt}, \tau_{\text{FG}}^{opt}\right) = \underset{(\tau_{\text{BG}}, \tau_{\text{FG}})}{\arg\max} \left(\underset{CDNet}{\text{mean}} \left(\text{F}^{BGS+SEM} \right) \right) \quad . \qquad (5)$$

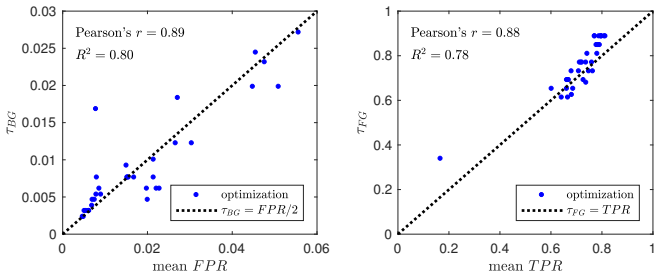

Fig. 2. Relationship between the thresholds $\left(\tau_{\text{BG}}^{opt}, \tau_{\text{FG}}^{opt}\right)$ defined by (5) and the mean performance of the BGS algorithm. The optimal thresholds are well approximated by the heuristic $\left(\tau_{\text{BG}}^{heu}, \tau_{\text{FG}}^{heu}\right) = \left(\text{FPR}^{BGS}/2, \text{TPR}^{BGS}\right)$.

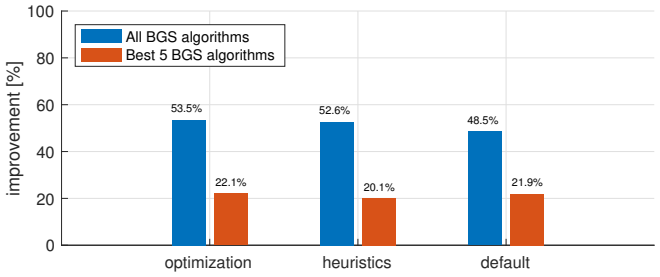

Fig. 3. Overall improvement, as defined in (4), for three parameter setting policies. For each policy, we manage to reduce significantly the overall error rate of BGS algorithms.

(2) Heuristics based policy. An analysis of these optimal thresholds showed that τ_{BG}^{opt} and τ_{FG}^{opt} are strongly correlated with FPR^{BGS} and TPR^{BGS} (see Figure 2). This led us to define the heuristics:

$$\left(\tau_{\text{BG}}^{heu}, \tau_{\text{FG}}^{heu}\right) = \left(\text{FPR}^{BGS}/2, \text{TPR}^{BGS}\right) \quad . \qquad (6)$$

These heuristics may be useful in practice for a BGS user who has access to the performance specifications of a BGS algorithm and hopes for good results without any time-consuming optimization process. Note that, as the BGS classifier performs better than a random classifier, we have $\text{FPR}^{BGS} < \text{TPR}^{BGS}$, which leads to $\tau_{\text{BG}}^{heu} < \tau_{\text{FG}}^{heu}$ given (6). The heuristics therefore guarantee that don't-care situations of Table 1 cannot be encountered.

(3) Default policy. A more simple alternative consists to set the pair $(\tau_{\text{BG}},\tau_{\text{FG}})$ to default values, such as the mean optimal thresholds of the 5 best BGS algorithms (according to the ranking of CDNet 2014), that is:

$$(\tau_{\text{BG}}, \tau_{\text{FG}}) = (0.00366,\ 0.88627). \qquad (7)$$

Figure 3 presents the improvement on the overall CDNet dataset for the three parameter setting policies. The three policies lead to very similar improvements and allow to reduce the mean overall ER of the best 5 BGS algorithms by more than 20%. Considering all BGS algorithms, we manage

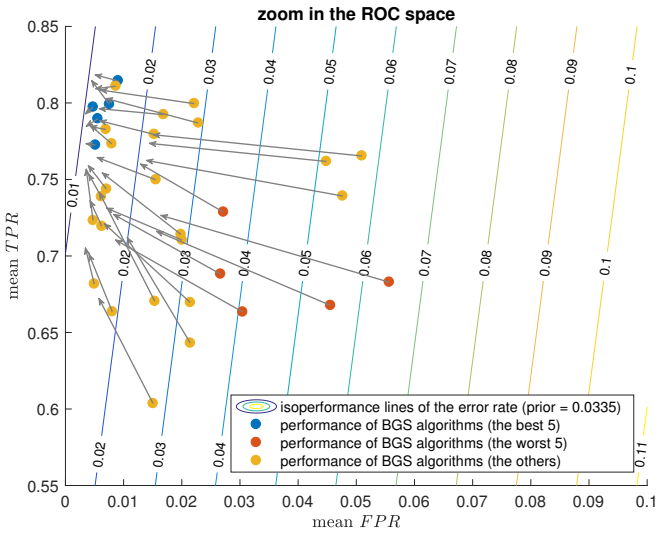

Fig. 4. Effect of our framework on the position of BGS classifiers in the overall ROC space of the CDNet dataset [16], with the default pair of thresholds given by (7). It tends to reduce the FPR significantly, while simultaneously increasing the TPR.

to reduce the mean overall ER by approximately 50%. Figure 4 shows that our framework tends to reduce significantly the FPR of BGS algorithms, while increasing simultaneously their TPR.

Per-category results are detailed in Figure 5. It turns out that our framework improves radically the segmentation masks on many categories (including "Dynamic background", "Shadow", and "PTZ"). Importantly, it should be noted that this does not come at the cost of deteriorating the results on the other categories.

Figure 6 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.

The consequence for these detection improvements is the computational overhead introduced by the semantic segmentation algorithm. The PSPNet model [19] used in our experiments runs at approximately 7 frames per second for 473×473 image resolution on a NVIDIA GeForce GTX Titan X GPU. However, it is possible to exploit the temporal stability of semantics in the video to reduce the computational load of the semantic segmentation, as done in [20]. Note that the computational load of (1), (2) and (3) is negligible compared to the computational load of semantic segmentation.

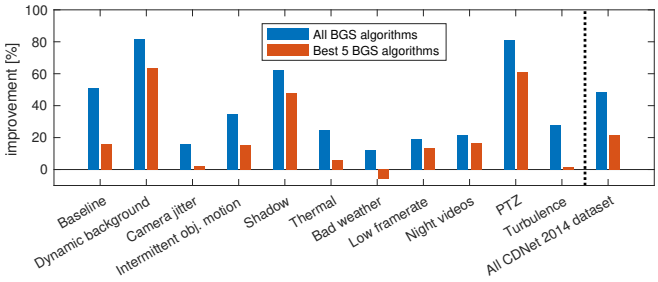

Fig. 5. Per-category mean improvements (see (4)) of our framework using the default pair of thresholds given by (7).

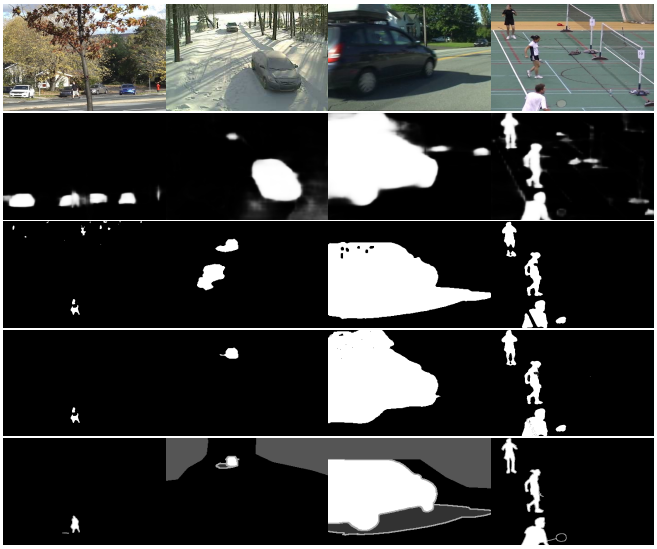

Fig. 6. Our framework addresses robustly dynamic backgrounds (column 1), ghosts (column 2) and strong shadows (column 3). In addition, it limits camouflage effects (column 4). From top row to bottom row: the input image, the probabilities $p_{S,t}(x)$, the output of IUTIS-5 [21], the output of IUTIS-5 integrated in our framework, and the ground truth.

4. 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 performances of 34 BGS algorithms, by reducing their mean overall error rate by roughly 50%.

Acknowledgments. Marc Braham has a grant funded by the FRIA. We are grateful to Hengshuang Zhao for publishing his PSPNet model [19] on the Internet. The GeForce GTX Titan X GPU used for this research was donated by the NVIDIA Corporation.

5. REFERENCES

[1] P.-M. Jodoin, S. Piérard, Y. Wang, and M. Van Droogenbroeck, "Overview and benchmarking of motion detection methods," in *Background Modeling and Foreground Detection for Video Surveillance*, chapter 24. Chapman and Hall/CRC, July 2014.

[2] T. Bouwmans, "Traditional and recent approaches in background modeling for foreground detection: An overview," *Computer Science Review*, vol. 11-12, pp. 31–66, May 2014.

[3] C. Stauffer and E. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Int. Conf. Comput. Vision and Pattern Recogn. (CVPR)*, June 1999, vol. 2, pp. 246–252.

[4] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1709–1724, June 2011.

[5] V. Jain, B. Kimia, and J. Mundy, "Background modeling based on subpixel edges," in *IEEE Int. Conf. Image Process. (ICIP)*, Sept. 2007, vol. 6, pp. 321–324.

[6] S. Zhang, H. Yao, and S. Liu, "Dynamic background modeling and subtraction using spatio-temporal local binary patterns," in *IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2008, pp. 1556–1559.

[7] M. Chen, Q. Yang, Q. Li, G. Wang, and M.-H. Yang, "Spatiotemporal background subtraction using minimum spanning tree and optical flow," in *Eur. Conf. Comput. Vision (ECCV)*. Sept. 2014, vol. 8695 of *Lecture Notes Comp. Sci.*, pp. 521–534, Springer.

[8] M. Braham, A. Lejeune, and M. Van Droogenbroeck, "A physically motivated pixel-based model for background subtraction in 3D images," in *IEEE Int. Conf. 3D Imaging (IC3D)*, Dec. 2014, pp. 1–8.

[9] T. Bouwmans, C. Silva, C. Marghes, M. Zitouni, H. Bhaskar, and C. Frelicot, "On the role and the importance of features for background modeling and foreground detection," *CoRR*, vol. abs/1611.09099, pp. 1–131, Nov. 2016.

[10] S. Piérard and M. Van Droogenbroeck, "A perfect estimation of a background image does not lead to a perfect background subtraction: analysis of the upper bound on the performance," in *Int. Conf. Image Anal. and Process. (ICIAP), Workshop Scene Background Modeling and Initialization (SBMI)*. Sept. 2015, vol. 9281 of *Lecture Notes Comp. Sci.*, pp. 527–534, Springer.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Int. Conf. Comput. Vision and Pattern Recogn. (CVPR)*, June 2015, pp. 3431–3440.

[12] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr, "Conditional random fields as recurrent neural networks," in *IEEE Int. Conf. Comput. Vision (ICCV)*, Dec. 2015, pp. 1529–1537.

[13] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Int. Conf. Comput. Vision and Pattern Recogn. (CVPR)*, July 2017.

[14] H. Zhao, "Pyramid scene parsing network," `https://www.youtube.com/watch?v=gdAVqJn_J2M`.

[15] L. Sevilla-Lara, D. Sun, V. Jampani, and M. J. Black, "Optical flow with semantic segmentation and localized layers," in *IEEE Int. Conf. Comput. Vision and Pattern Recogn. (CVPR)*, June 2016, pp. 3889–3898.

[16] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDnet 2014: An expanded change detection benchmark dataset," in *IEEE Int. Conf. Comput. Vision and Pattern Recogn. Workshops (CVPRW)*, June 2014, pp. 393–400.

[17] "Pascal VOC challenge performance evaluation and download server," `http://host.robots.ox.ac.uk:8080/leaderboard/displaylb.php?challengeid=11&compid=6`.

[18] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *IEEE Int. Conf. Comput. Vision and Pattern Recogn. (CVPR)*, July 2017.

[19] H. Zhao, "PSPNet50 ADE20K caffemodel," `https://drive.google.com/file/d/0BzaU285cX7TCN1R3QnUwQ0hoMTA/view`.

[20] E. Shelhamer, K. Rakelly, J. Hoffman, and T. Darrell, "Clockwork convnets for video semantic segmentation," in *Eur. Conf. Comput. Vision Workshops (ECCV Workshops)*. Oct. 2016, vol. 9915 of *Lecture Notes Comp. Sci.*, pp. 852–868, Springer.

[21] S. Bianco, G. Ciocca, and R. Schettini, "How far can you get by combining change detection algorithms?," *CoRR*, vol. abs/1505.02921, May 2015.

Erratum

This version of the paper slightly differs from the one published at ICIP 2017. Since then, we have discovered a problem in the code that was used to generate our results. This is a corrected version of the paper, including the updated results. The new results are even better than those that we had initially reported.