Towards Benchmarking Scene Background Initialization

Lucia Maddalena[1(✉)] and Alfredo Petrosino[2]

[1] Institute for High-Performance Computing and Networking, National Research Council, Naples, Italy
lucia.maddalena@cnr.it

[2] Department of Science and Technology, University of Naples Parthenope, Naples, Italy
alfredo.petrosino@uniparthenope.it

Abstract. Given a set of images of a scene taken at different times, the availability of an initial background model that describes the scene without foreground objects is the prerequisite for a wide range of applications, ranging from video surveillance to computational photography. Even though several methods have been proposed for scene background initialization, the lack of a common groundtruthed dataset and of a common set of metrics makes it difficult to compare their performance. To move first steps towards an easy and fair comparison of these methods, we assembled a dataset of sequences frequently adopted for background initialization, selected or created ground truths for quantitative evaluation through a selected suite of metrics, and compared results obtained by some existing methods, making all the material publicly available.

Keywords: Background initialization · Video analysis · Video surveillance

1 Introduction

The scene background modeling process is characterized by three main tasks: 1) *model representation*, that describes the kind of model used to represent the background; 2) *model initialization*, that regards the initialization of this model; and 3) *model update*, that concerns the mechanism used for adapting the model to background changes along the sequence. These tasks have been addressed by several methods, as acknowledged by several surveys (e.g., [2,4]). However, most of these methods focus on the representation and the update issues, whereas limited attention is given to the model initialization. The problem of scene background initialization is of interest for a very vast audience, due to its wide range of application areas. Indeed, the availability of an initial background model that describes the scene without foreground objects is the prerequisite, or at least can be of help, for many applications, including video surveillance, video segmentation, video compression, video inpainting, privacy protection for videos, and computational photography (see [6]).

V. Murino et al. (Eds.): ICIAP 2015 Workshops, LNCS 9281, pp. 469–476, 2015.
DOI: 10.1007/978-3-319-23222-5_57

We state the general problem of *background initialization*, also known as bootstrapping, background estimation, background reconstruction, initial background extraction, or background generation, as follows: *Given a set of images of a scene taken at different times, in which the background is occluded by any number of foreground objects, the aim is to determine a model describing the scene background with no foreground objects.*

Depending on the application, the set of images can consist of a subset of initial sequence frames adopted for background training (e.g., for video surveillance), a set of non-time sequence photographs (e.g., for computational photography), or the entire available sequence. In the following, this set of images will be generally referred to as the *bootstrap sequence*.

In order to move first steps towards an easy and fair comparison of existing and future background initialization methods, we assembled the SBI dataset, a set of sequences frequently adopted for background initialization, including ground truths for quantitative evaluation through a selected suite of metrics (made publicly available through http://sbmi2015.na.icar.cnr.it), and compared results obtained by some existing methods.

2 Sequences

The SBI dataset includes seven bootstrap sequences extracted by original publicly available sequences that are frequently used in the literature to evaluate background initialization algorithms. COST 211 (sequence *Hall&Monitor* can be found at http://www.ics.forth.gr/cvrl/demos/NEMESIS/hall_monitor.mpg), ATON (dataset available at http://cvrr.ucsd.edu/aton/shadow/index.html), and PBI (dataset available at http://www.diegm.uniud.it/fusiello/demo/bkg/). In Table 1 we report, for each sequence, the name, the dataset it belongs

Table 1. Information on sequences adopted for evaluation

Name	Dataset	Original frames	Used frames	Original Resolution	Final Resolution
Hall&Monitor	COST 211	0-299	4-299	352x240	352x240
HighwayI	ATON	0-439	0-439	320x240	320x240
HighwayII	ATON	0-499	0-499	320x240	320x240
CaVignal	PBI	0-257	0-257	200x136	200x136
Foliage	PBI	0-399	6-399	200x148	200x144
People&Foliage	PBI	0-349	0-340	320x240	320x240
Snellen	PBI	0-333	0-320	146x150	144x144

to, the number of available frames, the subset of the frames adopted for testing, the original and the final resolution. The subsets have been selected in order to avoid the inclusion into the testing sequences of *empty* frames (frames not including foreground objects), while the final resolution has been chosen in order

to avoid problems in the computation of boundary patches for block-based methods. The ground truths (GT) have been manually obtained by either choosing one of the sequence frames free of foreground objects (not included into the subsets of used frames) or by stitching together empty background regions from different sequence frames.

3 Metrics

The metrics adopted to evaluate the accuracy of the estimated background models have been chosen among those used in the literature for background estimation. They are described in Table 2, where GT (Ground Truth) is the image containing the *true* background, CB (Computed Background) is the estimated background image computed with one of the background initialization methods, L is the maximum number of grey levels, N is the number of image pixels, and MSE is the Mean Squared Error between GT and CB images. While the last metric is defined only for color images, all the other metrics are expressly defined for gray-scale images. In the case of color images, they are generally applied to either the gray-scale converted image or the luminance component Y of a color space such as YCbCr. The latter approach has been chosen for measurements reported in §4.

Table 2. Metrics adopted for evaluation

Name	Description	Range
AGE	Average Gray-level Error: Average of the gray-level absolute difference between GT and CB images.	[0, L-1]
EPs	Total number of Error Pixels: An *error pixel* is a pixel of CB whose value differs from the value of the corresponding pixel in GT by more than some threshold τ (in the experiments τ=20).	$[0, N]$
pEPs	Percentage of Error Pixels: EPs/N.	$[0, 1]$
CEPs	Total number of Clustered Error Pixels: A *clustered error pixel* is any error pixel whose 4-connected neighbors are also error pixels.	$[0, N]$
pCEPs	Percentage of Clustered Error Pixels: CEPs/N.	$[0, 1]$
PSNR	Peak-Signal-to-Noise-Ratio: $PSNR = 10 \cdot \log_{10}\left((L-1)^2/MSE\right)$.	
MS-SSIM	MultiScale Structural Similarity Index: Defined in [10].	$[0, 1]$
CQM	Color image Quality Measure: Defined in [11].	

4 Experimental Results and Comparisons

In this study, we compared the results on the SBI dataset of six background initialization methods, whose classifications according to [6] are summarized in Table 3. In the reported experiments, the temporal **Median** for the color bootstrap sequences is computed for each pixel as the one that minimizes

Table 3. Classifications of the compared methods

	pixel-level	region-level	hybrid	recursive	non-recursive	blind	selective
Temporal Statistics:							
Median [7]	X				X	X	
SC-SOBS [5]			X	X			X
Subintervals of Stable Intensity:							
WS2006 [9]	X				X		X
CA2008 [3]	X				X		X
Model Completion:							
RSL2011 [8]		X		X			X
Optimal Labeling:							
Photomontage [1]			X		X		X

the sum of L_∞ distances of the pixel from all the other pixels. The **SC-SOBS** [5] background estimate is obtained as the result of the initial training of the software SC-SOBS (publicly available in the download section of the CVPRLab at http://cvprlab.uniparthenope.it) using for all the sequences the same default parameter values. Once the neural background model is computed, the background estimate is extracted for each pixel by choosing the modeling weight vector that is closest to the ground truth. **WS2006** has been implemented based on [9], and parameter values have been chosen among those suggested by the authors and providing the best overall results. Results for **RSL2011** [8] have been obtained through the related software publicly available at http://arma.sourceforge.net/background_est/. Results for **Photomontage** [1] have been obtained through the related software publicly available at http://grail.cs.washington.edu/projects/photomontage/ choosing the maximum likelihood image objective as data term for achieving visual smoothness. Finally, results for **CA2008** [3] have been kindly provided by the authors.

Fig. 1 shows the background images obtained by the compared methods on the SBI dataset, while Table 4 reports accuracy results according to the metrics described in §3 (in boldface the best results for each metric and each sequence).

For sequence *Hall&Monitor*, we observe few differences in initializing the background in image regions where foreground objects are more persistent during the sequence. A man walking straight down the corridor occupies the same image region for more than 65% of the sequence frames, while the briefcase is left on the small table for the last 60%. WS2006, RSL2011, Photomontage, and CA2008 well handle the walking man, but only CA2008 does not include the abandoned briefcase into the background. This qualitative analysis is confirmed by accuracy results in terms of pEPs and pCEPs values reported in Table 4. Moreover, AGE values are quite low for all the compared methods, due to the reduced size of foreground objects as compared to the image size. However, the worst AGE values are achieved by RSL2011 and Photomontage, despite their quite good qualitative results. Finally, all the compared methods achieve

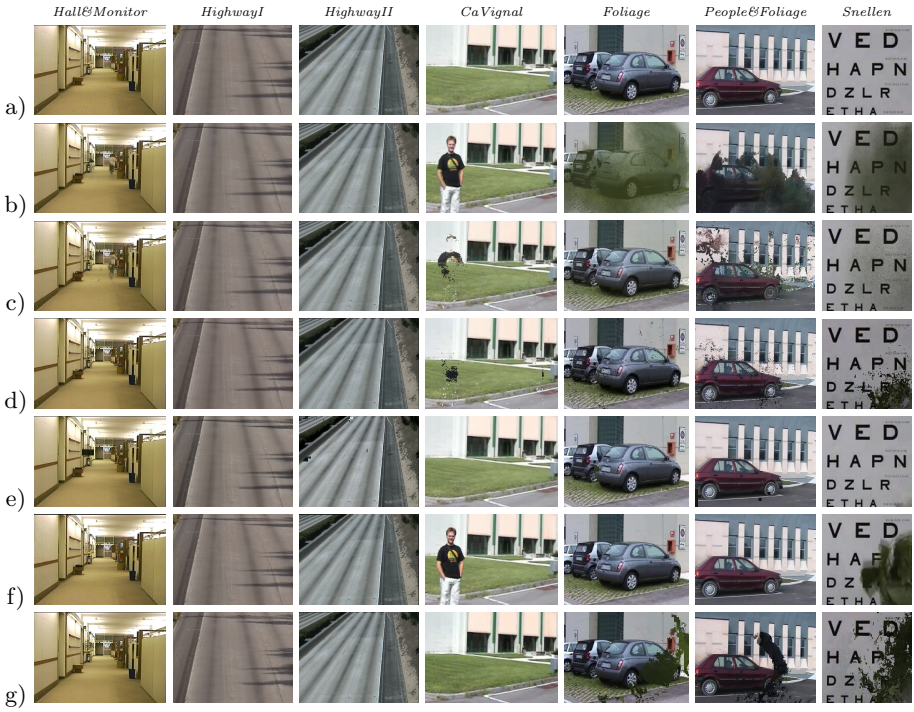

Fig. 1. Background initialization results on the SBI dataset obtained by: a) GT, b) Median, c) SC-SOBS, d) WS2006, e) RSL2011, f) Photomontage, g) CA2008

similar values of PSNR, MS-SSIM, and CQM, as overall, apart from reduced sized defects related to foreground objects, they all succeed in providing a sufficiently faithful representation of the empty background.

For both *HighwayI* and *HighwayII* sequences, all the compared methods succeed in providing an accurate estimated background. This is due to the fact that, even though the highway is always fairly crowded by passing cars, the background is revealed for at least 50% of the entire bootstrap sequence length and no cars remain stationary during the sequence. The above qualitative considerations are only partially confirmed by performance results reported in Table 4. Indeed, different AGE and pEPs values are achieved by qualitatively similar estimated backgrounds, while similar low pCEPs values and high MS-SSIM, PSNR, and CQM values are achieved by all the compared methods.

Sequence *CaVignal* represents a major burden for most of the compared methods. Indeed, the only man appearing in the sequence stands still on the left of the scene for the first 60% of sequence frames; then starts walking and rests on the right of the scene for the last 10% of sequence frames. The persistent clutter at the beginning of the scene leads most of the compared methods to include the man on the left into the estimated background, while the persistent clutter at the end of the scene leads only WS2006 to partially include the man on the right into

Table 4. Accuracy results of the compared methods on the SBI dataset

	Method	AGE	EPs	pEPs	CEPs	pCEPs	MS-SSIM	PSNR	CQM
Hall&Monitor	Median	2.7105	839	0.9931%	451	0.5339%	0.9640	30.4656	42.6705
	SC-SOBS	**2.4493**	828	0.9801%	272	0.3220%	0.9653	30.4384	**43.1867**
	WS2006	2.6644	470	0.5563%	26	0.0308%	0.9821	30.9313	40.0949
	RSL2011	3.2687	703	0.8321%	398	0.4711%	0.9584	28.4428	37.9971
	Photomontage	2.7986	**305**	**0.3610**%	69	0.0817%	0.9819	**33.3715**	41.7323
	CA2008	2.4737	337	0.3989%	**0**	**0.0000**%	**0.9878**	32.2503	41.2399
HighwayI	Median	1.4275	120	0.1563%	11	0.0143%	0.9924	40.1432	62.5723
	SC-SOBS	**1.2286**	**3**	**0.0039**%	**0**	**0.0000**%	**0.9949**	**42.6868**	**65.5755**
	WS2006	2.5185	526	0.6849%	19	0.0247%	0.9816	35.6885	56.9113
	RSL2011	2.8139	267	0.3477%	33	0.0430%	0.9830	36.0290	51.9835
	Photomontage	2.1745	313	0.4076%	37	0.0482%	0.9830	37.1250	59.0270
	CA2008	2.9477	895	1.1654%	65	0.0846%	0.9752	33.9800	56.1319
HighwayII	Median	1.7278	245	0.3190%	1	0.0013%	0.9961	34.6639	42.3162
	SC-SOBS	**0.6536**	**7**	**0.0091**%	**0**	**0.0000**%	**0.9982**	**44.6312**	**54.3785**
	WS2006	2.4906	375	0.4883%	10	0.0130%	0.9927	33.9515	40.5088
	RSL2011	5.6807	956	1.2448%	316	0.4115%	0.9766	28.6703	35.0821
	Photomontage	2.4306	452	0.5885%	4	0.0052%	0.9909	34.3975	41.7656
	CA2008	2.4340	486	0.6328%	43	0.0560%	0.9919	33.5545	39.4813
CaVignal	Median	10.3082	2846	10.4632%	2205	8.1066%	0.7984	18.1355	33.1438
	SC-SOBS	4.0941	869	3.1949%	436	1.6029%	0.8779	21.8507	42.2652
	WS2006	2.5403	408	1.5000%	129	0.4743%	0.9289	27.1089	37.0609
	RSL2011	**1.6132**	**4**	**0.0147**%	**0**	**0.0000**%	**0.9967**	**41.3795**	**52.5856**
	Photomontage	11.2665	3052	11.2206%	2408	8.8529%	0.7919	17.6257	32.0570
	CA2008	9.2569	17	0.0625%	**0**	**0.0000**%	0.9932	27.5197	39.7879
Foliage	Median	27.0135	13626	47.3125%	8772	30.4583%	0.6444	16.7842	28.7321
	SC-SOBS	3.8215	160	0.5556%	0	0.0000%	0.9900	31.7713	39.1387
	WS2006	6.8649	821	2.8507%	2	0.0069%	0.9754	27.2438	34.9776
	RSL2011	2.2773	43	0.1493%	11	0.0382%	0.9951	36.7450	43.1208
	Photomontage	**1.8592**	**0**	**0.0000**%	**0**	**0.0000**%	**0.9974**	**39.1779**	**45.6052**
	CA2008	18.3613	3327	11.5521%	1258	4.3681%	0.9092	18.7767	29.9137
People&Foliage	Median	24.4211	24760	32.2396%	19446	25.3203%	0.6114	15.1870	27.4979
	SC-SOBS	15.1031	10770	14.0234%	3849	5.0117%	0.7561	16.6189	35.3667
	WS2006	5.4243	2743	3.5716%	71	0.0924%	0.9269	22.6952	31.3847
	RSL2011	2.0980	612	0.7969%	434	0.5651%	0.9905	32.5550	37.0598
	Photomontage	**1.4103**	**3**	**0.0039**%	**0**	**0.0000**%	**0.9973**	**41.0866**	**47.1517**
	CA2008	19.7347	9401	12.2409%	4755	6.1914%	0.8220	17.1567	25.9970
Snellen	Median	42.3981	12898	62.2010%	11814	56.9734%	0.6932	13.6573	36.0691
	SC-SOBS	16.8898	7746	37.3553%	5055	24.3779%	0.9303	21.2571	44.7498
	WS2006	23.0010	4804	23.1674%	2544	12.2685%	0.7481	15.6158	24.9930
	RSL2011	**1.8095**	**133**	**0.6414**%	**99**	**0.4774**%	**0.9979**	**38.0295**	**50.2600**
	Photomontage	29.9797	6946	33.4973%	6318	30.4688%	0.5926	14.1466	26.9210
	CA2008	40.5218	9173	44.2371%	6359	30.6665%	0.6886	12.9428	24.0239

the background. RSL2011 and CA2008 perfectly handle the persistent clutter, even though only RSL2011 accordingly achieves the best accuracy results for all the metrics.

For sequence *Foliage*, even though moving leaves occupy most of the background area for most of the time, many of the compared methods achieve a quite good representation of the scene background. Indeed, only Median produces a greenish halo due to the foreground leaves over almost the entire scene area, and accordingly achieves the worst accuracy results for all the metrics.

In sequence *People&Foliage*, the artificially added leaves and men occupy almost all the scene area in almost all the sequence frames. Only Photomontage and RSL2011 appear to well handle the wide clutter, also achieving the best accuracy results for all the metrics.

In sequence *Snellen*, the foreground leaves occupy almost all the scene area in almost all the sequence frames. This leads most of the methods to include the contribution of leaves into the final background model. The best qualitative result can be attributed to RSL2011, as confirmed by the quantitative analysis in terms of all the adopted metrics.

Overall, we can observe that most of the best performing background initialization methods are region-based or hybrid, confirming the importance of taking into account spatio-temporal inter-pixel relations. Also selectivity in choosing the best candidate pixels, shared by all the best performing methods, appears to be important for achieving accurate results. Instead, all the common methodological schemes shared by the compared methods can lead to accurate results, showing no preferred scheme, and the same can be said concerning recursivity.

5 Concluding Remarks and Future Perspectives

We proposed a benchmarking study for scene background initialization, moving the first steps towards a fair and easy comparison of existing and future methods, on a common dataset of groundtruthed sequences, with a common set of metrics, and based on reproducible results. The assembled SBI dataset, the ground truths, and a tool to compute the suite of metrics were made publicly available.

Based on the benchmarking study, first considerations have been drawn. Concerning main issues in background initialization, low speed (or steadiness), rather than great size, of foreground objects included into the bootstrap sequence is a major burden for most of the methods. All the common methodologies shared by the compared methods can lead to accurate results, showing no preferred scheme, and the same can be said concerning recursivity. Anyway, the best results are generally achieved by methods that are region-based or hybrid, and selective; thus, these are the methods to be preferred. Concerning the evaluation of background initialization methods, among the eight selected metrics frequently adopted in the literature, pEPs and MS-SSIM confirm to be strongly indicative of the performance of background initialization methods.

Further insight will be certainly achieved by extending the dataset to include more sequences, also from other video categories (e.g., night videos and hardly

crowded scenes), the notion of ground truth, in order to better handle the case of multimodal backgrounds, and the evaluation metrics, also evaluating their robustness in handling different scene conditions and their combination to provide global evaluation scores.

Acknowledgments. This research was supported by Project PON01_01430 PT2LOG under the Research and Competitiveness PON, funded by the European Union (EU) via structural funds, with the responsibility of the Italian Ministry of Education, University, and Research (MIUR).

References

1. Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., Salesin, D., Cohen, M.: Interactive digital photomontage. ACM Trans. Graph. **23**(3), 294–302 (2004)
2. Bouwmans, T.: Traditional and recent approaches in background modeling for foreground detection: An overview. Computer Science Review **1112**, 31–66 (2014)
3. Chen, C.C., Aggarwal, J.: An adaptive background model initialization algorithm with objects moving at different depths. In: 15th IEEE International Conference on Image Processing, 2008. ICIP 2008, pp. 2664–2667 (2008)
4. Elhabian, S., El Sayed, K., Ahmed, S.: Moving object detection in spatial domain using background removal techniques: State-of-art. Recent Patents on Computer Science **1**(1), 32–54 (2008)
5. Maddalena, L., Petrosino, A.: The SOBS algorithm: what are the limits? In: Proc. CVPR Workshops, pp. 21–26, June 2012
6. Maddalena, L., Petrosino, A.: Background model initialization for static cameras. In: Bouwmans, T., Porikli, F., Hferlin, B., Vacavant, A. (eds.) Background Modeling and Foreground Detection for Video Surveillance, pp. 3-1-3-16. Chapman & Hall/CRC (2014)
7. Maddalena, L., Petrosino, A.: The 3dSOBS+ algorithm for moving object detection. Comput. Vis. Image Underst. **122**, 65–73 (2014)
8. Reddy, V., Sanderson, C., Lovell, B.C.: A low-complexity algorithm for static background estimation from cluttered image sequences in surveillance contexts. EURASIP J. Image Video Process. **2011**, 1:1–1:14 (2011)
9. Wang, H., Suter, D.: A novel robust statistical method for background initialization and visual surveillance. In: Narayanan, P.J., Nayar, S.K., Shum, H.-Y. (eds.) ACCV 2006. LNCS, vol. 3851, pp. 328–337. Springer, Heidelberg (2006)
10. Wang, Z., Simoncelli, E., Bovik, A.: Multiscale structural similarity for image quality assessment. In: Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2004, vol. 2, pp. 1398–1402 (2003)
11. Yalman, Y., Erturk, I.: A new color image quality measure based on YUV transformation and PSNR for human vision system. Turkish J. of Electrical Eng. & Comput. Sci. **21**, 603–612 (2013)