2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Extracting a Background Image by a Multi-modal Scene Background Model

Lucia Maddalena
Institute for High-Performance Computing and Networking
National Research Council
Naples, Italy
Email: lucia.maddalena@cnr.it

Alfredo Petrosino
Department of Science and Technology
University of Naples Parthenope
Naples, Italy
Email: alfredo.petrosino@uniparthenope.it

***Abstract*—In scene analysis, the availability of an initial background model that describes the scene without foreground objects is at the basis of many computer vision applications. Multi-modal models of the scene background are frequently adopted in the applications, where each mode tries to keep track of the multiple background modes observed along the sequence. In this work we specifically address the problem of extracting a single background image by a multi-modal model of the scene background, in order to compare it against a given ground truth image of the background. Experimental results are provided on the SBMnet dataset, based on an existing multi-model background model and different extraction criteria, and general conclusions are drawn.**

I. Introduction

The availability of an initial background model that describes a scene without foreground objects is at the basis of many applications, ranging from intelligent video surveillance to computational photography [1], [2], [3], [4]. Therefore, scene background initialization is a problem of interest for a very vast audience. Also known as bootstrapping, background estimation, background reconstruction, initial background extraction, or background generation, its aim is to determine an image describing the scene background empty of foreground objects, given a set of images of the scene taken at different times, in which the background is occluded by any number of foreground objects [5]. Depending on the application, the set of images (in the following referred to as the *bootstrap sequence*) 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.

The problem of background initialization has been afforded in several researches (for a survey of several methods, possible classifications, and general issues, the reader is referred to [5] and references therein; further recent works are also summarized in [6], [7]). Most of the proposed methods are specifically designed for background initialization, such as those relying on subintervals of stable intensity (e.g., [8]), iterative model completion (e.g., [9]), optimal labeling (e.g., [10]), or missing data reconstruction (e.g., [11]), eventually corroborated by foreground detection (e.g., [12]). They provide a uni-modal model of the scene background, i.e., a single image of the estimated background, and their evaluation using benchmarking datasets for background initialization, such as the SBI dataset [6], the dataset adopted in [7], and the SBMnet dataset (http://scenebackgroundmodeling.net), is carried out comparing such images with ground truth (GT) background images.

Any background modeling method, as those devised for the detection of foreground objects through background subtraction (see [13] for a recent survey), can also be clearly adopted for extracting an estimate of the scene background image. However, in the frequent case of multi-modal background models (e.g., MOG [14], KDE [15], and SOBS [16]), an issue not yet deeply investigated is how to automatically select, for each image pixel, the *mode* – among those stored in the model – that is best suited for being adopted in the scene background image. This issue naturally arises when evaluating multi-modal background subtraction methods for the purpose of background initialization [6], [17], [18], [19].

In this work we specifically address the problem of extracting a single background image by a multi-modal model of the scene background, proposing and analyzing different criteria. Rather than proposing a new background initialization algorithm, our aim is to investigate possible ways to exploit existing multi-modal background modeling algorithms for the background initialization problem.

In Section II we briefly describe the SC-SOBS background model [20], taken in our investigation as an example of multi-modal model for the scene background. Four different criteria, adapted to the SC-SOBS model, are proposed and justified in Section III. Experimental results using these criteria are provided in Section IV using the SBMnet dataset, and compared to the results of other background estimation methods. Section V summarizes the achieved conclusions.

II. The SC-SOBS Background Model

The background model constructed and maintained in the SC-SOBS algorithm [20] for background subtraction is based on the idea of building an image sequence neural background model by learning in a self-organizing manner image sequence variations, seen as trajectories of pixels in time. The network behaves as a competitive neural network that implements a winner-take-all function, with an associated mechanism that modifies the local synaptic plasticity of neurons, allowing learning to be spatially restricted to the local neighborhood

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

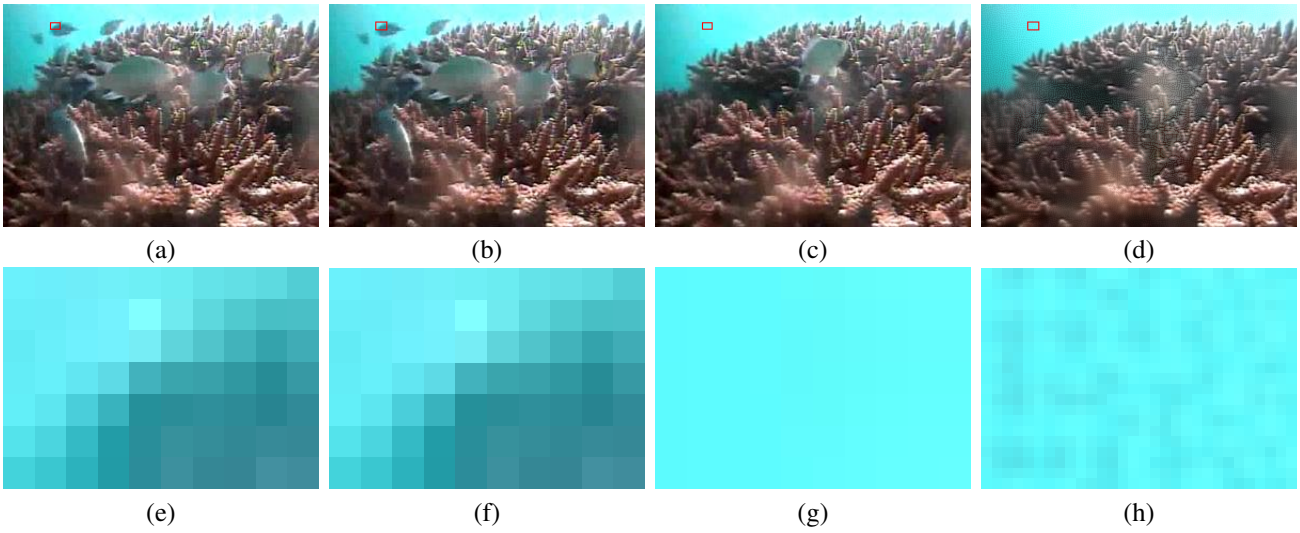

Fig. 1. Sequence *Hybrid* of the SBMnet dataset. First row: (a) frame I_0 (size 296×208); (b) SC-SOBS model B_0 (size 888×624); (c) frame I_{230}; (d) updated SC-SOBS model B_{230}. Second row: details (corresponding to the small red rectangle) of the same images of the first row (best viewed online).

of the most active neurons. Therefore, the neural background model well adapts to scene changes, capturing the most persisting features of the image sequence.

Given an image sequence $\{I_t\}$, for each pixel $\mathbf{p}$ in the image domain D, a neuronal map is built consisting of $n \times n$ weight vectors $m_t^{i,j}(\mathbf{p}), i, j$=0, …, n-1, which will be called a *model* for pixel $\mathbf{p}$ and will be indicated as $M_t(\mathbf{p})$:

$$M_t(\mathbf{p}) = \left\{ m_t^{i,j}(\mathbf{p}),\ i,j = 0,\ldots,n-1 \right\}. \qquad (1)$$

The complete set of models $M_t(\mathbf{p})$ for all pixels $\mathbf{p}$ of the t-th sequence frame I_t is organized as a 2D neuronal map B_t, where each pixel is represented by a local $n \times n$ neuronal map. This configuration allows us to easily take into account the spatial relationship among pixels and corresponding weight vectors.

For each pixel $\mathbf{p}$, the corresponding weight vectors of the model $M_0(\mathbf{p})$ are initialized with the pixel brightness value at time $t = 0$ ($m_0^{i,j}(\mathbf{p}) = I_0(\mathbf{p}),\ i,j = 0,\ldots,n-1$). Therefore, the resulting neuronal map B_0, obtained for all pixels $\mathbf{p}$, is an $n \times n$ enlarged version of the first sequence frame I_0. For example, the initial neuronal map B_0, shown in Fig. 1-(b) as an image of size 888×624, has been obtained by the first sequence frame I_0 of size 296×208 (Fig. 1-(a)) choosing $n = 3$, as in all the experiments reported in Section IV.

At each subsequent time step t, background subtraction is achieved by comparing each pixel $\mathbf{p}$ of the t-th sequence frame I_t with the current pixel model $M_{t-1}(\mathbf{p})$, in order to determine if there exists a best matching weight vector in $M_{t-1}(\mathbf{p})$ that is close enough to it. To this end, in the experiments reported in Section IV, the Euclidean distance of vectors in the HSV color hexcone has been adopted. If no acceptable matching weight vector exists, $\mathbf{p}$ is detected as a foreground pixel; otherwise, it means that $\mathbf{p}$ is a background pixel.

In case of spatially coherent background pixels, further learning of the neuronal map enables the adaptation of the background model to slight scene modifications. Such learning is achieved by updating the neural weights according to a visual attention mechanism of reinforcement, where the best matching weight vectors, together with their neighborhood, are reinforced into the neuronal map. As an example, in Fig. 1-(d) we report the updated SC-SOBS model B_{230} at time $t = 230$. Looking at the model detail shown in Fig. 1-(h), we can see that the updated background model stores many more variations of the blue color than the current sequence frame (Fig. 1-(g)). These are the updated weight vectors, that represent the most persisting background values observed along the sequence (mainly the water sky blue, but also darker blue values corresponding to swimming fish).

The above described initialization and update procedures are generally adopted for training, over the initial K training frames of a given sequence, the neural network background model, to be used for detection and adaptation in all subsequent sequence frames. What differentiates the training and the adaptation phases is the choice of method parameters. Specifically, during training, the foreground segmentation threshold is assigned a high value, so as to obtain a (possibly rough) initial background model that includes several observed pixel intensity variations, and the learning factor for model update is chosen as a monotonically decreasing function of time t, in order to ensure neural network convergence. For further details related to the background model update procedure and all the method parameters, the reader is referred to [20].

III. Extracting a Background Image

For the purpose of background initialization, the SC-SOBS background model B_K is computed as the result of the initial training over the entire bootstrap sequence $I_0, \ldots, I_K$, and, as outlined in Section II, it consists of n^2 weight vectors for each pixel. The background image estimate BE, of the same size of the bootstrap sequence, can be extracted according to

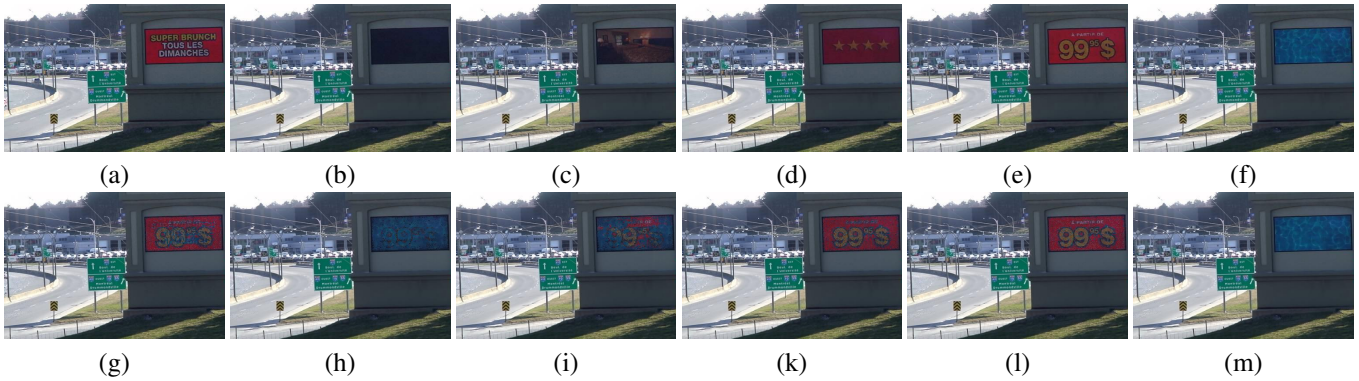

Fig. 2. Sequence *advertisementBoard* of the *backgroundMotion* category: (a)-(f) GT background images, representing 6 different contents of the changing advertisement board, shown in order of appearance; (g)-(m) results of SC-SOBS-C3 for each of the corresponding GT (best viewed online).

several procedures. In this paper, we considered four different criteria for each pixel **p**:

- **C1 (multi-modes average)**: as the mean of the model $M_K(\mathbf{p})$ for pixel **p**

$$BE(\mathbf{p}) = \text{mean}_{i,j=0,\ldots,n-1}(m_K^{i,j}(\mathbf{p})).$$

  This way of proceeding is analogous to the one adopted in publicly available software for extracting a single background image by a multi-modal background model (e.g., MOG2 in OpenCV [21]). However, it should be stressed that taking the average of the stored *modes* (the n^2 weight vectors for SC-SOBS or the k Gaussian means for MOG [14]) has little meaning as compared to the whole multi-modal model, that tries to keep track of the (possibly very different) background modes during the bootstrap sequence.
- **C2 (multi-modes vs. last frame)**: by choosing, in the model $M_K(\mathbf{p})$ for pixel **p**, the weight vector $BE(\mathbf{p})$ that is closest to the corresponding pixel in the last frame of the bootstrap sequence

$$d(BE(\mathbf{p}), I_K(\mathbf{p})) = \min_{i,j=0,\ldots,n-1} d(m_K^{i,j}(\mathbf{p}), I_K(\mathbf{p})), \quad (2)$$

  according to a distance $d(\cdot)$ (in the experiments it is the L_2 norm of the RGB values). This way of proceeding is strictly related to the motivation of any background subtraction method (i.e., change detection), since the model estimated during a bootstrap sequence is to be adopted for background subtraction with subsequent sequence frames (represented in this case by the last bootstrap frame I_K).
- **C3 (multi-modes vs. ground truth)**: by choosing the modeling weight vector that is closest to the ground truth image GT (as we did in [6]), i.e., substituting I_K with GT in Eq. (2). The aim here is to provide the best representation of the background that can be achieved through the constructed background multi-modal model, even though it is only applicable for comparison purposes, being based on the knowledge of a ground truth to compare with.
- **C4 (multi-modes vs. reference uni-mode)**: by choosing the modeling weight vector that is closest to the corresponding pixel in the background image R estimated by the most accurate uni-modal background initialization method, i.e., substituting I_K with R in Eq. (2). The aim here is to provide the best representation of the background that can be achieved through the constructed background multi-modal model, in a way that, at the same time, is independent on the knowledge of a ground truth background. As byproduct, it shows to what extent the multi-modal method outperforms the best uni-modal one.

We explicitly observe that the above reasonings, formulated for the SC-SOBS model adopted as an example, can be easily extended to many other multi-modal background subtraction methods.

IV. Experimental Results on SBMnet

Our experimental analysis is based on the SBMnet dataset, that includes 79 different videos, each with one or more GT background images, spanning 8 categories selected to include diverse scene background modeling challenges (Basic, Intermittent Motion, Clutter, Jitter, Illumination Changes, Background Motion, Very Long, and Very Short).

The metrics adopted for SBMnet have been chosen among those frequently used in the literature for background initialization, as described in detail in [6]. Lower values for AGE, pEPs, and pCEPs, as well as higher values for PSNR, MS-SSIM, and CQM, indicate higher accuracy of the estimated background image. For each sequence, they are evaluated comparing the estimated background images with a GT background image. For selected sequences, showing different conditions along the bootstrap sequences, more than one GT background image is available; in these cases, the reported performance results are those that maximize the MS-SSIM value, in accordance with the SBMnet evaluation scripts.

Background images extracted by the SC-SOBS model using the four procedures outlined in Section III have been compared to those obtained with other background initialization methods. As in [22], the temporal ColorMedian background estimate is computed, for each pixel, as the one that minimizes the sum

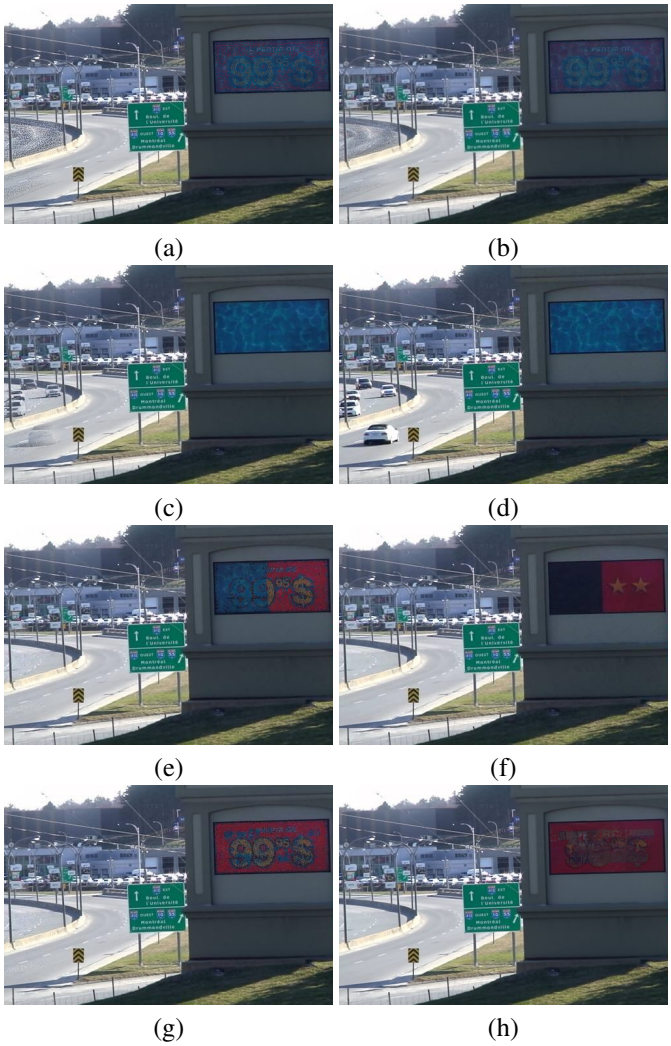

Fig. 3. Sequence *advertisementBoard* of the *backgroundMotion* category: (a) SC-SOBS model; (b) results of SC-SOBS-C1; (c) result of SC-SOBS-C2 (extracted using the last sequence frame shown in (d)); (e) result of SC-SOBS-C4 (extracted using as reference the result of Photomontage shown in (f)); (g) SC-SOBS-C4b (extracted using as reference the result of ColorMedian shown in (h)) (best viewed online).

of L_∞ distances of the RGB value from all RGB values for that pixel along the sequence. Results for Photomontage [10] have been obtained through the software made available by the authors, choosing the maximum likelihood image objective as data term for achieving visual smoothness. Being based on a batch processing algorithm, in the case of too long sequences, a subset of equidistant frames along the sequence has been selected. Finally, the background estimates extracted by the SC-SOBS model have been obtained as described in Section III. Specifically, the SC-SOBS model has been obtained as the result of the initial training of the related software (publicly available in the download section of the CVPRLab at http://cvprlab.uniparthenope.it) using for all the sequences the same default parameter values. Results indicated as SC-SOBS-C4 are obtained using criterion **C4**, choosing as reference background estimation method Photomontage, that, according to Table I, is the most accurate uni-modal background estimation method among those here considered for comparison. For further analysis, we also report results SC-SOBS-C4b, obtained by choosing the temporal Color Median, that, on average, is the second most accurate compared method. It should be made clear that the adopted ranking, that naturally arises from the average overall results of Table I, can clearly differ from the average ranking across categories reported on the SBMnet website.

Examples of background images extracted by the SC-SOBS model using the four procedures outlined in Section III and by the compared methods are reported in Figs. 2 and 3. In the first row of Fig. 2, we report the six GT images provided in the SBMnet dataset for sequence *advertisementBoard*, representing the scene empty of foreground objects, with 6 different contents of the changing advertisement board. In the second row of Fig. 2, we report the result of SC-SOBS-C3 for each of the corresponding GT. From these results we can observe that the background model (reported in Fig. 3-(a)), consisting in all the experiments of 3×3 weight vectors for each pixel, well represents the ground truth background appearing at the end of the video sequence, as shown in Fig. 2-(m) and, to some extent, in Fig. 2-(l). This is due to the SC-SOBS update strategy, that tends to incorporate into the background model the most recent background pixels. The mean adopted in SC-SOBS-C1, instead, produces an averaged version of the advertisement boards (shown in Fig. 3-(b)), corresponding to none of the GT background images. Using the last sequence frame (shown in Fig. 3-(d)), SC-SOBS-C2 produces an advertisement board (shown in Fig. 3-(c)) very similar to the last GT reported in Fig. 2-(f), but it also includes part of the moving cars along the highway, that have pixel colors similar to those stored in the SC-SOBS model. Using as reference the result of Photomontage (shown in Fig. 3-(f)), SC-SOBS-C4 produces an advertisement board closest to it (shown in Fig. 3-(e)), but again using the stored *modes* that better represent the last sequence frames. Finally, using as reference the result of temporal ColorMedian (shown in Fig. 3-(h)), SC-SOBS-C4b produces an advertisement board very similar to the GT reported in Fig. 2-(e).

Quantitative performance results on SBMnet are compared in Tables I and II in terms of average performance measures achieved by the different background initialization methods overall and on each video category, respectively. Here, we observe that performance measures achieved by SC-SOBS-C3 are the best for all categories. This shows the extreme ability of the SC-SOBS model to represent the scene background. However, as already pointed out in Section III, the way the estimated background images are extracted by the model according to the criterion **C3** is too *biased*, as compared to the other methods. This is the reason why we report in boldface the best results for each metric among all compared methods, but excluding SC-SOBS-C3.

Among results extracted by the SC-SOBS model, worst performance measures are those achieved by SC-SOBS-C1. This confirms the unsuitability of the usual practice for extracting a

TABLE I
AVERAGE ACCURACY RESULTS ON THE WHOLE SBMNET DATASET. IN BOLDFACE/UNDERLINED THE BEST/SECOND BEST RESULTS FOR EACH METRIC AMONG ALL COMPARED METHODS (EXCLUDING SC-SOBS-C3).

Method	AGE	pEPs	pCEPs	MS-SSIM	PSNR	CQM
ColorMedian	8.1150	0.0959	0.0526	0.9133	27.6703	28.6064
Photomontage	**7.1950**	**0.0686**	0.0257	**0.9189**	**28.0113**	**28.8719**
SC-SOBS-C1	10.6887	0.1499	0.0665	0.8843	24.8675	25.9016
SC-SOBS-C2	7.7161	0.0808	0.0335	0.9047	26.9576	27.9199
SC-SOBS-C4	7.5183	0.0711	**0.0242**	0.9160	27.6533	28.5601
SC-SOBS-C4b	7.9262	0.0776	0.0324	0.9086	27.6354	28.5362
SC-SOBS-C3	2.5863	0.0161	0.0045	0.9800	36.3332	36.8729

background image from a multi-modal background model by averaging its *modes*.

According to Table I, performance measures achieved by SC-SOBS-C4 are on average the best among those extracted by the SC-SOBS model. Even though the reference background estimation method (Photomontage) achieves overall results better than SC-SOBS-C4, there are many categories (namely, *Basic, IntermittentMotion, Jitter, BackgroundMotion*, and *VeryLong*) for which SC-SOBS-C4 results are better (see Table II). The reason is that the multi-modal SC-SOBS model is able to store different background values, chosen among the most persistent along the bootstrap sequence, and thus excluding eventual non persistent foreground values. For the *IlluminationChanges* category, instead, SC-SOBS-C4 achieves much worse performance results than Photomontage. Example results for sequence *CameraParameter* from this category are reported in Fig. 4 and Table III. For this sequence, two GT images are provided in SBMnet, representing the scene with lights off and on (Figs. 4-(a)-(b)). The result of Photomontage (Fig. 4-(d)) is close to the more frequent dark scenario, while the SC-SOBS model (Fig. 4-(c)), being updated for online background subtraction, tends to better represent the lit scenario of the last bootstrap frames. Therefore, SC-SOBS-C4 (Fig. 4-(e)), extracted taking for reference the dark scenario, achieves poor performance results. Better results, in this case, are achieved taking into account the bright last sequence frame (see result of SC-SOBS-C2 in Fig. 4-(f)).

Further considerations can be drawn comparing SC-SOBS-C4 and SC-SOBS-C4b results, obtained using the **C4** criterion and different reference uni-modal background estimation methods. Generally, as expected, better background images are extracted by taking as reference the more accurate method, as shown in Table I, even though there may be isolated cases not fulfilling this expectation (e.g., for the VeryShort category, where slightly better results are achieved by SC-SOBS-C4b as compared to those of SC-SOBS-C4).

V. CONCLUSION

Driven by the need to compare different background models for the purpose of background estimation, in this paper we analyzed different automatic criteria for extracting a single estimated background image by a multi-modal model, exemplifying several alternatives based on the SC-SOBS multi-modal background model.

TABLE II
ACCURACY RESULTS ON ALL CATEGORIES OF THE SBMNET DATASET. IN BOLDFACE THE BEST RESULTS FOR EACH METRIC AMONG ALL COMPARED METHODS (EXCLUDING SC-SOBS-C3).

	Method	AGE	pEPs	pCEPs	MS-SSIM	PSNR	CQM
Basic	ColorMedian	**3.6941**	**0.0127**	0.0035	**0.9810**	**34.1813**	**34.8247**
	Photomontage	4.4856	0.0226	0.0039	0.9719	32.3208	32.9621
	SC-SOBS-C1	8.5898	0.1125	0.0475	0.9287	26.4904	27.5031
	SC-SOBS-C2	5.6780	0.0418	0.0179	0.9466	28.5685	29.5205
	SC-SOBS-C4	4.3598	0.0200	0.0033	0.9728	32.1766	32.8665
	SC-SOBS-C4b	3.9135	0.0146	**0.0033**	0.9777	32.9309	33.6122
	SC-SOBS-C3	1.8025	0.0036	0.0017	0.9902	39.4453	39.8313
Int.Motion	ColorMedian	6.8457	0.0612	0.0414	0.9151	24.7058	25.7739
	Photomontage	7.1460	0.0639	0.0427	0.9138	24.8941	25.8682
	SC-SOBS-C1	9.5116	0.1323	0.0551	0.8976	24.6332	25.7379
	SC-SOBS-C2	6.2452	0.0518	0.0290	0.9396	26.6425	27.7076
	SC-SOBS-C4	6.2583	0.0487	**0.0238**	0.9255	25.9249	26.9569
	SC-SOBS-C4b	6.1912	0.0486	0.0241	0.9251	25.7893	26.8305
	SC-SOBS-C3	2.6263	0.0137	0.0078	0.9781	34.1659	34.9035
Clutter	ColorMedian	12.4760	0.1555	0.1066	0.8120	26.0386	27.1007
	Photomontage	**6.8195**	**0.0543**	**0.0294**	0.8892	**28.5554**	**29.4882**
	SC-SOBS-C1	18.0379	0.2982	0.1890	0.7959	21.8034	23.0811
	SC-SOBS-C2	12.9403	0.1640	0.0915	0.7900	21.5548	22.8110
	SC-SOBS-C4	7.0590	0.0644	0.0304	**0.8939**	28.0077	29.0737
	SC-SOBS-C4b	12.7395	0.1568	0.1034	0.8102	25.4252	26.5450
	SC-SOBS-C3	3.6306	0.0294	0.0113	0.9653	33.5012	34.3982
Jitter	ColorMedian	**8.9660**	**0.1048**	**0.0401**	**0.8565**	**25.6888**	**26.7869**
	Photomontage	10.1272	0.1210	0.0441	0.8390	24.3478	25.4186
	SC-SOBS-C1	12.9314	0.1990	0.0812	0.7905	22.3104	23.4806
	SC-SOBS-C2	10.7645	0.1315	0.0559	0.8179	23.8171	24.9466
	SC-SOBS-C4	10.0232	0.1186	0.0420	0.8403	24.5562	25.6570
	SC-SOBS-C4b	9.3232	0.1078	0.0398	0.8506	25.2905	26.3732
	SC-SOBS-C3	3.8419	0.0355	0.0090	0.9577	32.9675	33.7220
Ill. Changes	ColorMedian	12.0055	0.2308	0.1768	0.9377	24.3424	25.4479
	Photomontage	5.2668	0.0329	0.0155	0.9743	30.2102	31.0393
	SC-SOBS-C1	11.0782	0.1404	0.0729	0.9242	24.7584	25.6874
	SC-SOBS-C2	**4.7231**	**0.0300**	**0.0139**	**0.9827**	**30.6951**	**31.4548**
	SC-SOBS-C4	10.3591	0.1005	0.0574	0.9075	26.2190	27.0837
	SC-SOBS-C4b	10.9937	0.1081	0.0571	0.9031	25.2372	26.1682
	SC-SOBS-C3	1.7964	0.0074	0.0009	0.9940	38.1984	38.6331
Bckg. Motion	ColorMedian	**9.0640**	**0.1200**	0.0253	0.8679	**26.3857**	**27.3097**
	Photomontage	12.0930	0.1589	0.0410	0.8244	23.5420	24.5253
	SC-SOBS-C1	10.5269	0.1516	0.0330	0.8626	24.7998	25.7293
	SC-SOBS-C2	10.1155	0.1373	0.0329	0.8627	25.4669	26.4329
	SC-SOBS-C4	10.7280	0.1481	0.0302	0.8486	24.5806	25.5603
	SC-SOBS-C4b	9.2921	0.1236	**0.0240**	**0.8692**	25.8423	26.7056
	SC-SOBS-C3	3.4983	0.0308	0.0027	0.9738	34.2457	34.6913
VeryLong	ColorMedian	6.8762	0.0549	0.0208	**0.9848**	29.5425	30.4075
	Photomontage	6.6446	0.0629	0.0259	0.9838	29.2081	30.0166
	SC-SOBS-C1	6.7501	0.0645	0.0106	0.9683	28.0284	28.9539
	SC-SOBS-C2	**4.2929**	**0.0280**	0.0051	0.9714	**31.1008**	**31.8405**
	SC-SOBS-C4	6.0638	0.0355	0.0021	0.9837	29.2615	30.1014
	SC-SOBS-C4b	5.6626	0.0308	**0.0015**	0.9847	30.4258	31.1957
	SC-SOBS-C3	1.2410	0.0008	0.0001	0.9963	41.3309	41.6148
VeryShort	ColorMedian	4.9923	0.0277	0.0060	0.9515	30.4774	31.1999
	Photomontage	**4.9770**	0.0327	**0.0030**	0.9548	**31.0117**	**31.6568**
	SC-SOBS-C1	8.0835	0.1003	0.0424	0.9070	26.1158	27.0397
	SC-SOBS-C2	6.9668	0.0616	0.0257	0.9266	27.8168	28.6472
	SC-SOBS-C4	5.2953	0.0330	0.0044	**0.9556**	30.4997	31.1813
	SC-SOBS-C4b	5.2764	**0.0309**	0.0073	0.9493	30.2797	30.9889
	SC-SOBS-C3	2.2531	0.0077	0.0024	0.9848	36.8109	37.1887

TABLE III
ACCURACY RESULTS ON *CameraParameter* SEQUENCE.

Method	AGE	pEPs	pCEPs	MS-SSIM	PSNR	CQM
Photomontage	1.6291	0.0003	0.0000	0.9925	40.4538	40.7403
SC-SOBS-C2	7.4852	0.0688	0.0265	0.9796	23.7660	24.7126
SC-SOBS-C4	30.9091	0.3493	0.2219	0.6086	13.4793	14.7779

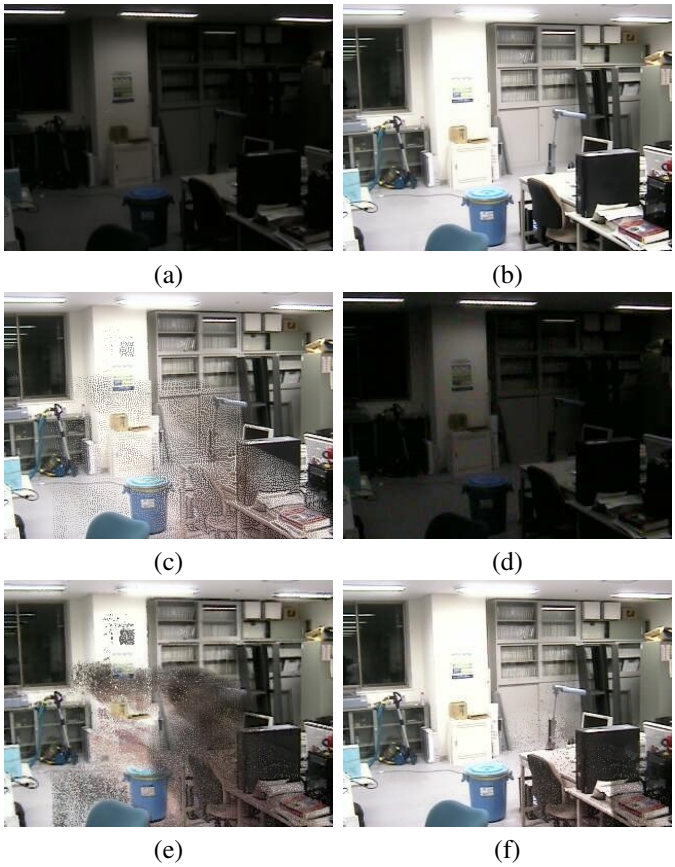

Fig. 4. Sequence *CameraParameter* of the *illuminationChanges* category: (a) and (b) GT background images, representing the scene with lights off and on, respectively; (c) SC-SOBS model; results of (d) Photomontage, (e) SC-SOBS-C4, and (f) SC-SOBS-C2.

Following extensive experimental results carried out on the SBMnet dataset (http://scenebackgroundmodeling.net), we showed the unsuitability of the usual practice for extracting a background image from a multi-modal background model by averaging its *modes*. Instead, the best criterion resulted the one that, for each pixel, considers the *mode* that is closest to the corresponding pixel in the background image computed with the most accurate uni-modal background estimation algorithm. The choice for such reference estimated background image can easily be done looking at the benchmarking results available through the SMBnet website.

ACKNOWLEDGMENT

The authors would like to thank Prof. Pierre-Marc Jodoin, Université de Sherbrooke, Canada, for useful discussions on the subject. This research was supported by Project PON01_01430 PT2LOG, funded by the European Union and MIUR and by LAB GPT Project, funded by MIUR.

REFERENCES

[1] L. Maddalena and A. Petrosino, "Object motion detection and tracking by an artificial intelligence approach," *Intern. J. of Pattern Recognition and Artificial Intelligence*, vol. 22, no. 5, pp. 915–928, Jan. 2008.

[2] R. Melfi, S. Kondra, and A. Petrosino, "Human activity modeling by spatio temporal textural appearance," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1990–1994, 2013.

[3] L. Maddalena, A. Petrosino, and F. Russo, "People counting by learning their appearance in a multi-view camera environment," *Pattern Recognition Letters*, vol. 36, pp. 125–134, 2014.

[4] M. E. Maresca and A. Petrosino, "Clustering local motion estimates for robust and efficient object tracking," in *Computer Vision - ECCV 2014 Workshops*, L. Agapito, M. M. Bronstein, and C. Rother, Eds. Springer International Publishing, 2015, pp. 244–253.

[5] L. Maddalena and A. Petrosino, "Background model initialization for static cameras," in *Background Modeling and Foreground Detection for Video Surveillance*, T. Bouwmans, F. Porikli, B. Hoferlin, and A. Vacavant, Eds. Chapman & Hall/CRC, 2014, pp. 3–1–3–16.

[6] ——, "Towards benchmarking scene background initialization," in *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*, V. Murino et al., Ed. Springer, 2015, pp. 469–476.

[7] D. Ortego, J. C. SanMiguel, and J. M. Martnez, "Rejection based multipath reconstruction for background estimation in video sequences with stationary objects," *Comput. Vis. Image Underst.*, vol. 147, pp. 23–37, 2016.

[8] C.-C. Chen and J. Aggarwal, "An adaptive background model initialization algorithm with objects moving at different depths," in *Proc. ICIP*, 2008, pp. 2664–2667.

[9] V. Reddy, C. Sanderson, and B. C. Lovell, "A low-complexity algorithm for static background estimation from cluttered image sequences in surveillance contexts," *EURASIP J. Image Video Process.*, vol. 2011, pp. 1:1–1:14, Jan. 2011.

[10] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen, "Interactive digital photomontage," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 294–302, Aug. 2004.

[11] A. Sobral, T. Bouwmans, and E.-h. Zahzah, "Comparison of matrix completion algorithms for background initialization in videos," in *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*, V. Murino et al., Ed. Springer, 2015, pp. 510–518.

[12] B. Laugraud, S. Piérard, M. Braham, and M. Van Droogenbroeck, "Simple median-based method for stationary background generation using background subtraction algorithms," in *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*, V. Murino et al., Ed. Springer, 2015, pp. 477–484.

[13] A. Sobral and A. Vacavant, "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos," *Comput. Vis. Image Underst.*, vol. 122, pp. 4–21, 2014.

[14] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. CVPR*, vol. 2, 1999, p. 252.

[15] A. M. Elgammal, D. Harwood, and L. S. Davis, "Non-parametric model for background subtraction," in *Proc. ECCV*. London, UK: Springer-Verlag, 2000, pp. 751–767.

[16] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1168–1177, July 2008.

[17] D. D. Bloisi, A. Grillo, A. Pennisi, L. Iocchi, and C. Passaretti, "Multi-modal background model initialization," in *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*, V. Murino et al., Ed. Springer, 2015, pp. 485–492.

[18] M. De Gregorio and M. Giordano, "Background modeling by weightless neural networks," in *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*, V. Murino et al., Ed. Springer, 2015, pp. 493–501.

[19] N. Noceti, A. Staglianò, A. Verri, and F. Odone, "Bmtdl for scene modeling on the sbi dataset," in *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*, V. Murino et al., Ed. Springer, 2015, pp. 502–509.

[20] L. Maddalena and A. Petrosino, "The SOBS algorithm: What are the limits?" in *Proc. CVPR Workshops*, June 2012, pp. 21–26.

[21] G. Bradski, *Dr. Dobb's Journal of Software Tools*, 2000.

[22] L. Maddalena and A. Petrosino, "The 3dSOBS+ algorithm for moving object detection," *Comput. Vis. Image Underst.*, vol. 122, pp. 65–73, 2014.