A Benchmarking Framework for Background Subtraction in RGBD Videos

Massimo Camplani[1], Lucia Maddalena[2(✉)], Gabriel Moyá Alcover[3], Alfredo Petrosino[4], and Luis Salgado[5,6]

[1] University of Bristol, Bristol, UK
[2] National Research Council, Naples, Italy
lucia.maddalena@cnr.it
[3] Universitat de les Illes Balears, Palma, Spain
[4] University of Naples Parthenope, Naples, Italy
[5] Universidad Politécnica de Madrid, Madrid, Spain
[6] Universidad Autónoma de Madrid, Madrid, Spain

Abstract. The complementary nature of color and depth synchronized information acquired by low cost RGBD sensors poses new challenges and design opportunities in several applications and research areas. Here, we focus on background subtraction for moving object detection, which is the building block for many computer vision applications, being the first relevant step for subsequent recognition, classification, and activity analysis tasks. The aim of this paper is to describe a novel benchmarking framework that we set up and made publicly available in order to evaluate and compare scene background modeling methods for moving object detection on RGBD videos. The proposed framework involves the largest RGBD video dataset ever made for this specific purpose. The 33 videos span seven categories, selected to include diverse scene background modeling challenges for moving object detection. Seven evaluation metrics, chosen among the most widely used, are adopted to evaluate the results against a wide set of pixel-wise ground truths. Moreover, we present a preliminary analysis of results, devoted to assess to what extent the various background modeling challenges pose troubles to background subtraction methods exploiting color and depth information.

Keywords: Background subtraction · Color and depth data · RGBD

1 Introduction

The advent of low cost RGBD sensors such as Microsoft Kinect or Asus Xtion Pro is completely changing the computer vision world, as they are being successfully used in several applications and research areas. Many of these applications, such as gaming or human computer interaction systems, rely on the efficiency of learning a scene background model for detecting and tracking moving objects, to be further processed and analyzed. Depth data is particularly attractive and suitable for applications based on moving objects detection, since

S. Battiato et al. (Eds.): ICIAP 2017 International Workshops, LNCS 10590, pp. 219–229, 2017.
https://doi.org/10.1007/978-3-319-70742-6_21

they are not affected by several problems representative of color-based imagery. However, depth data suffer from other problems, such as depth camouflage or depth sensor noisy measurements, which limit the efficiency of depth-only based background modeling approaches. The complementary nature of color and depth synchronized information acquired by RGBD sensors poses new challenges and design opportunities. New strategies are required that explore the effectiveness of the combination of depth- and color-based features, or their joint incorporation into well known moving object detection and tracking frameworks.

In order to evaluate and compare scene background modelling methods for moving object detection on RGBD videos, we assembled and made available the SBM-RGBD dataset[1]. It provides all facilities (data, ground truths, and evaluation scripts) for the SBM-RGBD Challenge, organized in conjunction with the Workshop on Background Learning for Detection and Tracking from RGBD Videos, 2017. The dataset and the results of the SBM-RGBD Challenge, which are described in the following sections, will remain available also after the competition, as reference for future methods.

2 Video Categories

The SBM-RGBD dataset provides a wide set of synchronized color and depth sequences acquired by the Microsoft Kinect. The dataset consists of 33 videos (about 15000 frames) representative of typical indoor visual data captured in video surveillance and smart environment scenarios, selected to cover a wide range of scene background modeling challenges for moving object detection. The videos come from our personal collections as well as from existing public datasets, including the GSM dataset, described in Moyá-Alcover et al. [13], MULTIVISION, described in Fernandez-Sanchez et al. [5], the Princeton Tracking Benchmark, described by Song and Xiao [14], the RGB-D object detection dataset, described by Camplani and Salgado [3], and the UR Fall Detection Dataset, described by Kwolek and Kepski [7].

The videos have 640×480 spatial resolution and their length varies from 70 to 1400 frames. Depth images are recorded at either 16 or 8 bits. They are already synchronized and registered with the corresponding color images by projecting the depth map onto the color image, allowing a color-depth pixel correspondence. For each sequence, pixels that have no color-depth correspondence (due to the difference in the color and depth cameras centers) are indicated in black in a binary Region-of-Interest (ROI) image (see Fig. 2-(c)) and are excluded by the evaluation (see Sect. 4).

The videos span seven categories, selected to include diverse scene background modelling challenges for moving object detection. These well known challenges can be related only to the RGB channels (RGB), only to the depth channel (D), or can be related to all the channels (RGB+D):

[1] http://rgbd2017.na.icar.cnr.it/SBM-RGBDdataset.html.

1. **Bootstrapping** (RGB+D): Videos including foreground objects in all their frames. The challenge is to learn a model of the scene background (to be adopted for background subtraction) even when the usual assumption of having a set of training frames empty of foreground objects fails.
   This category includes five videos, in most of which the background is never shown in some scene regions, being always occupied by foreground people.
2. **Color Camouflage** (RGB): Videos including foreground objects whose color is very close to that of the background, making hard a correct segmentation based only on color.
   This category consists of four videos where foreground objects are moved in front of similarly colored background (e.g., a white box in front of other white boxes or a rolling furniture moving in front of other furniture of the same color).
3. **Depth Camouflage** (D): Videos including foreground objects very close in depth to the background. Indeed, in these cases the sensor gives the same depth data values for foreground and background, making hard a correct segmentation based only on depth.
   The category consists of four videos where people move their hands or other objects very close to the background.
4. **Illumination Changes** (RGB): Videos containing strong and mild illumination changes. The challenge here is to adapt the color background model to illumination changes in order to achieve an accurate foreground detection. Four videos are included into this category, where the illumination varies due to the covering of the light source or to unstable illumination acquisition.
5. **Intermittent Motion** (RGB+D): Videos with scenarios known for causing ghosting artifacts in the detected motion, i.e., abandoned foreground objects or removed foreground objects. The challenge here is to detect foreground objects even if they stop moving (abandoned object) or if they were initially stationary and then start moving (removed object).
   This category consists of six videos including abandoned and removed objects. Two videos are obtained by reversing the original temporal order of the frames (so that an object that is abandoned in the original sequence results as removed in the reversed sequence).
6. **Out of Sensor Range** (D): Videos including foreground or background objects that are too close to/far from the sensor. Indeed, in these cases the sensor is unable to measure depth, due to its minimum and maximum depth specifications, resulting in *invalid* depth values.
   Five videos are included into this category, where several invalid depth values are due to foreground objects whose distance from the sensor is out of the admissible sensor range.
7. **Shadows** (RGB+D): Videos showing shadows caused by foreground objects. Indeed, foreground objects block the active light emitted by the sensor from reaching the background. This causes the casting on the background of shadows, that apparently behave as moving objects. RGBD sensors exhibit two different types of shadows: visible-light shadows in the RGB channels or IR shadows in the depth channel.

Fig. 1. Examples of videos from all the categories: (a) Bootstrapping, (b) ColorCamouflage, (c) DepthCamouflage, (d) IlluminationChanges, (e) IntermittentMotion, (f) OutOfRange, (g) Shadows.

The category consists of five videos including more or less strong shadows.

Examples of videos from all the categories are reported in Fig. 1.

3 Ground Truths

To enable a precise quantitative comparison of various algorithms for moving object detection from RGBD videos, all the videos come with pixel-wise ground truth foreground segmentations for each video. A foreground region is intended as anything that does not belong to the background, including abandoned objects and still persons, but excluding light reflections, shadows, etc. The ground truth images, some of which created using the GroundTruther software kindly made available by the organizers of changedetection.net, contain four labels (see Fig. 2-(d)), namely:

- 0: Background
- 85: Outside ROI
- 170: Unknown motion
- 255: Foreground

Areas around moving objects are labeled as *unknown motion*, due to semi-transparency and motion blur that do not allow a precise foreground/background classification. Therefore, these areas, as those not included into the ROI, are excluded by the evaluation.

While our evaluation is made across all the ground truths for all the videos, only a subset of the available ground truths is made publicly available for testing, in order to reduce the possibility of overtuning method parameters.

4 Metrics

The SBM-RGBD dataset comes also with tools to compute performance metrics for moving object detection from RGBD videos, and thus identify algorithms that are robust across various challenges. Let TP, FP, FN, and TN indicate,

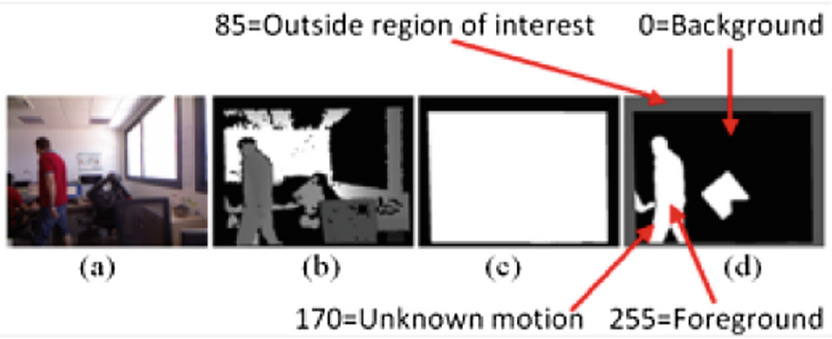

Fig. 2. Sequence ChairBox: (a) color and (b) depth images; (c) ROI; (d) ground truth.

for each video, the total number of True Positive, False Positive, False Negative, and True Negative pixels, respectively. The seven adopted metrics, widely adopted in the literature for evaluating the results of moving object detection (e.g., [6]), are

1. Recall
$$Rec = \frac{TP}{TP + FN}$$
2. Specificity
$$Sp = \frac{TN}{TN + FP}$$
3. False Positive Rate
$$FPR = \frac{FP}{FP + TN}$$
4. False Negative Rate
$$FNR = \frac{FN}{TP + FN}$$
5. Percentage of Wrong Classifications
$$PWC = 100 * \frac{FN + FP}{TP + FN + FP + TN}$$
6. Precision
$$Prec = \frac{TP}{TP + FP}$$
7. F-Measure
$$F_1 = \frac{2 * Prec * Rec}{Prec + Rec}$$

The Matlab scripts to compute all performance metrics have been adapted by the scripts available from changedetection.net.

5 Experimental Results

Several authors submitted their results to the SBM-RGBD challenge, and some of them provided a description of their method: RGBD-SOBS and RGB-SOBS [11], SCAD [12], and cwisardH+ [4]. Therefore, our experimental analysis is mainly devoted to assess to what extent the different background modelling challenges introduced in Sect. 2 pose troubles to these background subtraction methods.

In Table 1, we report average results on the whole dataset achieved by all submitted methods (as of July 4th, 2017), while in Tables 2 and 3, we report their average results for each challenge category[2].

Table 1. Average results on the whole SBM-RGBD dataset.

Method Name	*Rec*	*Sp*	*FPR*	*FNR*	*PWC*	*Prec*	F_1
RGBD-SOBS [11]	0.8391	0.9958	0.0042	0.0895	1.0828	0.8796	0.8557
RGB-SOBS [11]	0.7707	0.9708	0.0292	0.1578	5.4010	0.7247	0.7068
SRPCA [2]	0.7786	0.9739	0.0261	0.1499	3.1911	0.7474	0.7472
AvgM-D	0.7065	0.9869	0.0131	0.2221	2.8848	0.7498	0.7157
Kim	0.8493	0.9947	0.0053	0.0793	1.0292	0.8764	0.8606
SCAD [12]	0.8847	0.9932	0.0068	0.0439	0.9088	0.8698	0.8757
cwisardH+ [4]	0.7622	0.9817	0.0183	0.1664	2.8806	0.7556	0.7470

Bootstrapping can be a problem, especially for selective background subtraction methods (e.g., [9]), i.e. those that update the background model using only background information. Indeed, once a foreground object is erroneously included into the background model (e.g., due to inappropriate background initialization or to inaccurate segmentation of foreground objects), it will hardly be removed by the model, continuing to produce false negative results. The problem is even harder if some parts of the background are never shown during the sequences, as it happens in most of the videos of the Bootstrapping category. Indeed, in these cases, also the best performing background initialization methods [1] fail, as illustrated in Fig. 3, and only alternative techniques (e.g., inpainting) can be adopted to recover missing data [10]. Nonetheless, depth information seems to be beneficial for affording the challenge, as reported in Table 2, where accurate results are achieved by most of the methods that exploit depth information.

As expected, all the methods that exploit depth information achieve high accuracy in case of *color camouflage*. An evident example of the benefits induced by depth information for this category is given by the F-measure value achieved

[2] All the results are available at http://rgbd2017.na.icar.cnr.it/SBM-RGBDchallenge Results.html.

Table 2. Average results for each category of the SBM-RGBD dataset (Part 1).

Method Name	*Rec*	*Sp*	*FPR*	*FNR*	*PWC*	*Prec*	F_1
Bootstrapping							
RGBD-SOBS	0.8842	0.9925	0.0075	0.1158	2.3270	0.9080	0.8917
RGB-SOBS	0.8023	0.9814	0.0186	0.1977	4.4221	0.8165	0.8007
SRPCA	0.7284	0.9914	0.0086	0.2716	3.7409	0.9164	0.8098
AvgM-D	0.4587	0.9861	0.0139	0.5413	7.1960	0.6941	0.5350
Kim	0.8805	0.9965	0.0035	0.1195	1.5227	0.9566	0.9169
SCAD	0.8997	0.9940	0.0060	0.1003	1.8015	0.9319	0.9134
cwisardH+	0.5727	0.9616	0.0384	0.4273	8.1381	0.5787	0.5669
ColorCamouflage							
RGBD-SOBS	0.9563	0.9927	0.0073	0.0437	1.2161	0.9434	0.9488
RGB-SOBS	0.4310	0.9767	0.0233	0.5690	16.0404	0.8018	0.4864
SRPCA	0.8476	0.9389	0.0611	0.1524	4.3124	0.8367	0.8329
AvgM-D	0.9001	0.9793	0.0207	0.0999	2.0719	0.8096	0.8508
Kim	0.9737	0.9927	0.0073	0.0263	0.7389	0.9754	0.9745
SCAD	0.9875	0.9904	0.0096	0.0125	0.7037	0.9677	0.9775
cwisardH+	0.9533	0.9849	0.0151	0.0467	1.1931	0.9502	0.9510
DepthCamouflage							
RGBD-SOBS	0.8401	0.9985	0.0015	0.1599	0.9778	0.9682	0.8936
RGB-SOBS	0.9725	0.9856	0.0144	0.0275	1.5809	0.8354	0.8935
SRPCA	0.8679	0.9778	0.0222	0.1321	2.9944	0.7850	0.8083
AvgM-D	0.8368	0.9922	0.0078	0.1632	1.6943	0.8860	0.8538
Kim	0.8702	0.9968	0.0032	0.1298	0.9820	0.9433	0.9009
SCAD	0.9841	0.9963	0.0037	0.0159	0.4432	0.9447	0.9638
cwisardH+	0.6821	0.9949	0.0051	0.3179	2.4049	0.9016	0.7648
IlluminationChanges							
RGBD-SOBS	0.4514	0.9955	0.0045	0.0486	0.9321	0.4737	0.4597
RGB-SOBS	0.4366	0.9715	0.0285	0.0634	3.5022	0.4759	0.4527
SRPCA	0.4795	0.9816	0.0184	0.0205	1.9171	0.4159	0.4454
AvgM-D	0.3392	0.9858	0.0142	0.1608	3.0717	0.4188	0.3569
Kim	0.4479	0.9935	0.0065	0.0521	1.1395	0.4587	0.4499
SCAD	0.4699	0.9927	0.0073	0.0301	0.9715	0.4567	0.4610
cwisardH+	0.4707	0.9914	0.0086	0.0293	1.0754	0.4504	0.4581
IntermittentMotion							
RGBD-SOBS	0.8921	0.9970	0.0030	0.1079	0.8648	0.9544	0.9202
RGB-SOBS	0.9265	0.9028	0.0972	0.0735	9.3877	0.4054	0.5397
SRPCA	0.8893	0.9629	0.0371	0.1107	3.7026	0.7208	0.7735
AvgM-D	0.8976	0.9912	0.0088	0.1024	1.4603	0.9115	0.9027
Kim	0.9418	0.9938	0.0062	0.0582	0.9213	0.9385	0.9390
SCAD	0.9563	0.9914	0.0086	0.0437	0.8616	0.9243	0.9375
cwisardH+	0.8086	0.9558	0.0442	0.1914	5.0851	0.5984	0.6633

Table 3. Average results for each category of the SBM-RGBD dataset (Part 2).

Method Name	*Rec*	*Sp*	*FPR*	*FNR*	*PWC*	*Prec*	F_1
OutOfRange							
RGBD-SOBS	0.9170	0.9975	0.0025	0.0830	0.5613	0.9362	0.9260
RGB-SOBS	0.8902	0.9896	0.0104	0.1098	1.3610	0.8237	0.8527
SRPCA	0.8785	0.9878	0.0122	0.1215	1.6100	0.7443	0.8011
AvgM-D	0.6319	0.9860	0.0140	0.3681	2.7663	0.6360	0.6325
Kim	0.9040	0.9961	0.0039	0.0960	0.8228	0.9216	0.9120
SCAD	0.9286	0.9965	0.0035	0.0714	0.5711	0.9357	0.9309
cwisardH+	0.8959	0.9956	0.0044	0.1041	0.8731	0.9038	0.8987
Shadows							
RGBD-SOBS	0.9323	0.9970	0.0030	0.0677	0.7001	0.9733	0.9500
RGB-SOBS	0.9359	0.9881	0.0119	0.0641	1.5128	0.9140	0.9218
SRPCA	0.7592	0.9768	0.0232	0.2408	4.0602	0.8128	0.7591
AvgM-D	0.8812	0.9876	0.0124	0.1188	1.9330	0.8927	0.8784
Kim	0.9270	0.9934	0.0066	0.0730	1.0771	0.9404	0.9314
SCAD	0.9665	0.9910	0.0090	0.0335	1.0093	0.9276	0.9458
cwisardH+	0.9518	0.9877	0.0123	0.0482	1.3942	0.9062	0.9264

by the RGBD-SOBS method, that doubles the value achieved by the same method but without considering depth (RGB-SOBS). A similar reasoning can be applied to the *illumination changes* challenge. However, we point out that, in this case, the analysis should be based on Specificity, FPR, FNR, and PWC, rather than on the other three metrics. Indeed, two of the four videos of this category have no foreground objects throughout the whole duration, their rationale being the willingness of not detecting false positives under varying illumination conditions. This leads to have no positive cases in all ground truths and, consequently, to undefined values of Precision, Recall, and F-measure (in the experiments, values for these undefined cases are set to zero).

Depth can be beneficial also for detecting and properly handling cases of *intermittent motion*. Indeed, foreground objects can be easily identified based on their depth, that is lower than that of the background, even when they remain stationary for long time periods. Methods that explicitly exploit this characteristic (e.g., RGBD-SOBS and SCAD) succeed in handling cases of removed and abandoned objects, achieving high accuracy.

Overall, *shadows* do not seem to pose a strong challenge to most of the methods. Indeed, depth shadows due to moving objects cause some undefined depth values, generally close to the object contours, but these can be handled based on motion. Color shadows can be handled either exploiting depth information, that is insensitive to this challenge, or through color shadow detection techniques (e.g., as in RGB-SOBS and SCAD), when only color information is taken into

Fig. 3. Background image for sequence adl24cam0 (where the center area of the room is always covered by the man) computed using: (a) temporal median filter and (b) LabGen [8].

account. Instead, they are still a challenge when the sole grey level intensity is considered (e.g., as in SRPCA).

Out of range and *Depth camouflage* are among the most challenging issues, at least when information on color is disregarded or not properly combined with depth. Indeed, even though accuracy of most of the methods is moderately high, several false negatives are produced, as shown in Fig. 4 for depth camouflage.

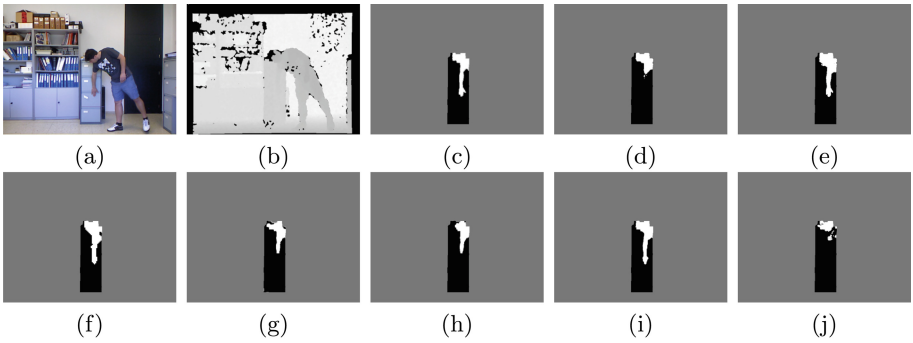

Fig. 4. Sequence DCamSeq2 (DepthCamouflage): (a) image no. 534, corresponding (b) depth image, and (c) ground truth; segmentation masks achieved by: (d) RGBD-SOBS, (e) RGB-SOBS, (f) SRPCA, (g) AvgM-D, (h) Kim, (i) SCAD, (j) CwisardH+.

6 Conclusions and Perspectives

The paper describes a novel benchmarking framework that we set up and made publicly available in order to evaluate and compare scene background modeling methods for moving object detection on RGBD videos. The SBM-RGBD

dataset is the largest RGBD video collection ever made available for this specific purpose. The 33 videos span seven categories, selected to include diverse scene background modeling challenges for moving object detection. Seven evaluation metrics, chosen among the most widely used, are adopted to evaluate the results against a wide set of pixel-wise ground truths. A preliminary analysis of results achieved by several methods investigates to what extent the various background modeling challenges pose troubles to background subtraction methods that exploit color and depth information. The proposed framework will serve as a reference for future methods aiming at overcoming these challenges.

Acknowledgments. We would like to thank all the authors who submitted their results to the SBM-RGBD Challenge, which will serve as reference for future generation methods. L. Maddalena wishes to acknowledge the GNCS (Gruppo Nazionale di Calcolo Scientifico) and the INTEROMICS Flagship Project funded by MIUR, Italy. A. Petrosino wishes to acknowledge Project VIRTUALOG Horizon 2020-PON 2014/2020. L. Salgado wishes to acknowledge projects TEC2013-48453 (MR-UHDTV) and TEC2016-75981 (IVME) funded by the Ministerio de Economa, Industria y Competitividad (AEI/FEDER) of the Spanish Government.

References

1. Bouwmans, T., Maddalena, L., Petrosino, A.: Scene background initialization: a taxonomy. Pattern Recogn. Lett. **96**, 3–11 (2017)
2. Bouwmans, T., Sobral, A., Javed, S., Jung, S.K., Zahzah, E.-H.: Decomposition into low-rank plus additive matrices for background/foreground separation: a review for a comparative evaluation with a large-scale dataset. Comput. Sci. Rev. **23**, 1–71 (2017)
3. Camplani, M., Salgado, L.: Background foreground segmentation with RGB-D Kinect data: an efficient combination of classifiers. J. Vis. Commun. Image Represent. **25**(1), 122–136 (2014)
4. De Gregorio, M., Giordano, M.: CwisarDH$^+$: Background Detection in RGBD Videos by Learning. In: Battiato, S., Gallo, G., Farinella, G., Leo, M. (eds.) ICIAP 2017. LNCS, vol. 10590, pp. 242–253. Springer, Cham (2017)
5. Fernandez-Sanchez, E.J., Diaz, J., Ros, E.: Background subtraction based on color and depth using active sensors. Sensors **13**, 8895–8915 (2013)
6. Goyette, N., Jodoin, P.M., Porikli, F., Konrad, J., Ishwar, P.: Changedetection.net: a new change detection benchmark dataset. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2012, pp. 1–8, June 2012
7. Kwolek, B., Kepski, M.: Human fall detection on embedded platform using depth maps and wireless accelerometer. Comput. Methods Programs Biomed. **117**(3), 489–501 (2014)
8. Laugraud, B., Piérard, S., Braham, M., Van Droogenbroeck, M.: Simple median-based method for stationary background generation using background subtraction algorithms. In: Murino, V., Puppo, E., Sona, D., Cristani, M., Sansone, C. (eds.) ICIAP 2015. LNCS, vol. 9281, pp. 477–484. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23222-5_58

9. Maddalena, L., Petrosino, A.: A self-organizing approach to background subtraction for visual surveillance applications. IEEE Trans. Image Process. **17**(7), 1168–1177 (2008)
10. Maddalena, L., Petrosino, A.: Background model initialization for static cameras. In: Bouwmans, T., Porikli, F., Hoferlin, B., Vacavant, A. (eds.) Background Modeling and Foreground Detection for Video Surveillance, pp. 3-1-3-16. Chapman & Hall/CRC (2014)
11. Maddalena, L., Petrosino, A.: Exploiting color and depth for background subtraction. In: Battiato, S., Gallo, G., Farinella, G., Leo, M. (eds.) ICIAP 2017. LNCS, vol. 10590, pp. 254–265. Springer, Cham (2017)
12. Minematsu, T., Shimada, A., Uchiyama, H., Taniguchi, R.: Simple combination of appearance and depth for foreground segmentation. In: Battiato, S., Gallo, G., Farinella, G., Leo, M. (eds.) ICIAP 2017. LNCS, vol. 10590, pp. 266–277. Springer, Cham (2017)
13. Moyá-Alcover, G., Elgammal, A., Jaume-i-Capó, A., Varona, J.: Modeling depth for nonparametric foreground segmentation using RGBD devices. Pattern Recogn. Lett. **96**, 76–85 (2017)
14. Song, S., Xiao, J.: Tracking revisited using RGBD camera: unified benchmark and baselines. In: Proceedings of the 2013 IEEE International Conference on Computer Vision, ICCV 2013, pp. 233–240. IEEE Computer Society (2013)