Journal of *Imaging*

MDPI

Article

Background Subtraction for Moving Object Detection in RGBD Data: A Survey

Lucia Maddalena [1,*] and Alfredo Petrosino [2]

[1] National Research Council, Institute for High-Performance Computing and Networking, 80131 Naples, Italy
[2] Department of Science and Technology, University of Naples Parthenope, 80143 Naples, Italy; alfredo.petrosino@uniparthenope.it
* Correspondence: lucia.maddalena@cnr.it; Tel.: +39-081-6139522

Received: 16 April 2018; Accepted: 9 May 2018; Published: 16 May 2018

Abstract: The paper provides a specific perspective view on background subtraction for moving object detection, as a building block for many computer vision applications, being the first relevant step for subsequent recognition, classification, and activity analysis tasks. Since color information is not sufficient for dealing with problems like light switches or local gradual changes of illumination, shadows cast by the foreground objects, and color camouflage, new information needs to be caught to deal with these issues. Depth synchronized information acquired by low-cost RGBD sensors is considered in this paper to give evidence about which issues can be solved, but also to highlight new challenges and design opportunities in several applications and research areas.

Keywords: background subtraction; color and depth data; RGBD

1. Introduction

Background modeling is a critical component for motion detection tasks, and it is essential for most of modern video surveillance applications. Usually, the color information provides most of the information useful to detect foreground and to solve all the basic issues related to this task [1–5]. Anyway, problems like light switches or local gradual changes of illumination, shadows cast by the foreground objects, and color camouflage due to similar color of foreground and background regions are still open. The recent broad availability of depth data (from stereo vision to off-the-shelf RGBD sensors, such as time-of-flight and structured light cameras) opened new ways of dealing with the problem. Indeed, dense depth data provided by RGBD cameras is very attractive for foreground/background segmentation in indoor environments (due to the range camera limitations), since it does not suffer from the above-mentioned challenging issues that affect color-based algorithms. Moreover, depth information is beneficial to detect and reduce the effection of moved background objects.

On the other hand, the use of just depth data poses several problems that do not assure the required efficiency: (a) depth-based segmentation fails in case of depth camouflage that appears when foreground objects move towards the modeled background; (b) object silhouettes are strongly affected by the high level of depth data noise at object boundaries; (c) depth measurements are not always available for all the image pixels due to multiple reflections, scattering in particular surfaces, or occlusions. All these issues arose with several background modeling approaches based solely on depth as proposed in [6–10], mainly as building blocks for people-detection and tracking systems [11–14].

Therefore, many recent methods try to exploit the complementary nature of color and depth information acquired with RGBD sensors. Generally, these methods either extend to RGBD data's well-known background models initially designed for color data [15,16] or model the scene background (and sometimes also the foreground) based on color and depth independently and then combine the results, on the basis of different criteria [17–20] (see Section 3).

Several reviews related to RGBD data have been recently presented. In [21], Cruz et al. provide one of the first surveys of academic and industrial research on Kinect and RGBD data, showing the basic principles to begin developing applications using Kinect. Greff et al. [8] present a comparison between background subtraction algorithms using depth cameras. In [22], Zhang unravels the intelligent technologies encoded in Kinect, such as sensor calibration, human skeletal tracking, and facial-expression tracking. It also demonstrates a prototype system that employs multiple Kinects in an immersive teleconferencing application. In [23], Han et al. present a comprehensive review of recent Kinect-based computer vision algorithms and applications, giving insights on how researchers exploit and improve computer vision algorithms using Kinect. In [24], Camplani et al. survey multiple human tracking in RGBD data.

The present paper aims to provide a comprehensive review of methods which exploit RGBD data for moving object detection based on background subtraction. We do not review methods based only on RGB features, as that would need a dedicated survey of its own and would demand much greater space—for RGB only background subtraction, the reader is referred to the reviews presented in [1–3,5]. We provide a brief analysis of the main issues and a concise description of the existing literature. Moreover, we summarize the metrics commonly used for the evaluation of these methods and the datasets that are publicly available. Finally, we provide the most extensive comparison of the existing methods on some datasets.

2. RGBD Data and Related Issues for Background Subtraction

Color cameras are based on sensors like CCD or CMOS, which provide a reliable representation of the scene with high-resolution images. Background subtraction using this kind of sensors often results in a precise separation between foreground and background, even though well-known scene background modeling challenges for moving object detection must be taken into account [25,26]:

- **Bootstrapping**: 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.
- **Color Camouflage**: When videos include foreground objects whose color is very close to that of the background, it is hard to provide a correct segmentation based only on color.
- **Illumination Changes**: The challenge is to adapt the color background model to strong or mild illumination changes to achieve an accurate foreground detection.
- **Intermittent Motion**: The issue is to detect foreground objects even if they stop moving (abandoned objects) or if they were initially stationary and then start moving (removed objects).
- **Moving Background**: The challenge is to model not only the static background but also slight changes in the background that are not interesting for surveillance, such as waving trees in outdoor videos.
- **Color Shadows**: The challenge is to discriminate foreground objects by shadows cast on the background by foreground objects that apparently behave as moving objects.

Depth sensors provide partial geometrical information about the scene that can help solving some of the above problems. A depth image, storing for each pixel a depth value proportional to the estimated distance from the device to the corresponding point in the real world, can be obtained with different methods [27]:

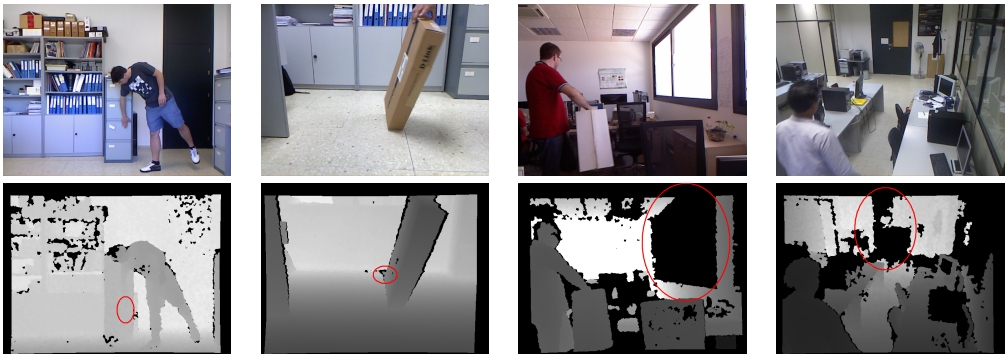
1. *Stereo vision* [28]: this is a passive technique where the depth is derived from the disparity between images captured from a camera pair. Stereo vision systems need to be well-calibrated and can fail when the scene is not sufficiently textured. Moreover, algorithms for stereo reconstruction are often computationally expensive. Finally, stereo vision systems cannot work in low light conditions. In this case, infrared (IR) lights can be added to the system, but then, the color information is lost, which generates segmentation and matching difficulties.

2. *Time-of-Flight* (ToF) [29]: ToF cameras are active sensors that determine the per-pixel depth value by measuring the time taken by IR light to travel to the object and back to the camera. A ToF camera provides more accurate depth images than a stereo vision system, but it is very expensive and limited to low image resolution. The measured depth map can be noisy both spatially and temporally, and noise is content-dependent and hence, difficult to remove by traditional filtering methods.
3. *Structured light* [30]: A structured light sensor consists of an IR emitter and an IR camera. The emitter projects an IR speckle pattern onto the scene; the camera captures the reflected pattern and correlates it against a stored reference pattern on a plane, providing the depth values. Well known examples include the Microsoft Kinect version 1 (in the following simply named Kinect) and the Asus Xtion Pro Live. These sensors can acquire higher resolution images than a ToF camera at a lower price. The drawback is that depth information is not always well estimated at the object boundaries and for areas too far from/too close to the IR projector. Also, the noise in depth measurements increases quadratically with increasing distance from the sensor [31].

Even though depth data solves some of the previously highlighted background maintenance issues, being independent of scene color and illumination conditions, it suffers from several problems, independent of which technology is used for its estimation. Indeed, as for color data, depth data suffers from bootstrapping, intermittent motion, and moving background. Moreover, challenges specific for depth data include [32,33].

- **Depth Camouflage**: When foreground objects are very close in depth to the background, the sensor gives the same depth data values for foreground and background, making it hard to provide a correct segmentation based only on depth.
- **Depth Shadows**: Similar to the case of color, depth shadows are caused by foreground objects blocking the IR light emitted by the sensor from reaching the background.
- **Specular Materials**: When the scene includes specular objects, IR rays from a single incoming direction are reflected back in a single outgoing direction, without causing the diffusion needed to obtain depth information.
- **Out of Sensor Range**: When foreground or background objects are too close to/far from the sensor, the sensor is unable to measure depth, due to its minimum and maximum depth specifications.

In the last three cases, where depth cannot be measured at a given pixel, the sensor returns a special non-value code to indicate its inability to measure depth [32], resulting in an *invalid* depth value (shown as black pixels in the depth images reported in Figure 1). These invalid values must be suitably handled to exploit depth for background subtraction.



(a) (b) (c) (d)

Figure 1. Background modeling issues related to depth data (highlighted by red ellipses). (**a**) Depth camouflage; (**b**) Depth shadows; (**c**) Specular materials; (**d**) Out of sensor range.

3. Methods

In the last twenty years, several methods have been proposed for background subtraction exploiting depth data, as an alternative or complement to color data. A summary of background subtraction methods for RGBD videos is given in Table 1. Here, apart from the name of the authors and the related reference (column Authors and Ref.), we report (column Used data) whether they exploit only the depth information (D) or the complementary nature of color and depth information (RGBD). Moreover, we specify (column Depth data) how the considered depth data is acquired (Kinect, ToF cameras, stereo vision devices). Furthermore, we specify (column Model) the type of model adopted for the background, including Codebook [34], Frame difference, Kernel Density Estimation (KDE) [35], Mixture of Gaussians (MoG) [36], Robust Principal Components Analysis (RPCA) [37], Self-Organizing Background Subtraction (SOBS) [38], Single Gaussian [39], Thresholding, ViBe [40], and WiSARD weightless neural network [41]. Finally, we specify (column No. of models) if they extend to RGBD data well-known background models originally designed for color data (1 model) or model the scene background based on color and depth independently and then combine the results, on the basis of different criteria (2 models).

In the following, we provide a brief description of the reviewed methods, presented in chronological order. In case of research dealing with higher-level systems (e.g., teleconferencing, matting, fall detection, human tracking, gesture recognition, object detection), we limit our attention to background modeling and foreground detection.

Eveland et al. [6] present a method of statistical background modeling for stereo sequences based on the disparity images extracted from stereo pairs. The depth background is modeled by a single Gaussian, similarly to [39], but selective update prevents the incorporation of foreground objects into the background.

The method proposed by Gordon et al. [42] is an adaptation of the MoG algorithm to color and depth data obtained with a stereo device. Each background pixel is modeled as a mixture of four-dimensional Gaussian distributions: three components are the color data (the YUV color space components), and the fourth one is the depth data. Color and depth features are considered independent, and the same updating strategy of the original MoG algorithm is used to update the distribution parameters. The authors propose a strategy where, for reliable depth data, depth-based decisions bias the color-based ones: in case that a reliable distribution match is found in the depth component, the color-based matching criterion is relaxed, thus reducing the color camouflage errors. When the stereo matching algorithm is not reliable, the color-based matching criterion is set to be harder to avoid problems such as shadows or local illumination changes.

Ivanov et al. [43] propose an approach based on stereo vision, which uses the disparity (estimated offline) to warp one image of the pair in the other one, thus creating a geometric background model. If the color and brightness between corresponding points do not match, the pixels either belong to a foreground object or to an occlusion shadow. The latter case can be further disambiguated using more than two camera views.

Harville et al. [16] propose a foreground segmentation method using the YUV color space with the additional depth values estimated by stereo cameras. They adopt four-dimensional MoG models, also modulating the background model learning rate based on scene activity and making color-based segmentation criteria dependent on depth observations.

Kolmogorov et al. [44] describe two algorithms for bi-layer segmentation fusing stereo and color/contrast information, focused on live background substitution for teleconferencing. To segment the foreground, this approach relies on stereo vision, assuming that people participating in the teleconference are close to the camera. Color information is used to cope with stereo occlusion and low-texture regions. The color/contrast model is composed of MoG models for the background and the foreground.

Table 1. Summary of background subtraction methods for RGBD videos.

Authors & Ref.	Used Data	Depth Data	Model	No. of Models
Eveland et al. (1998) [6]	D	Stereo	Single Gaussian	1
Gordon et al. (1999) [42]	RGBD	Stereo	MoG	1
Ivanov et al. (2000) [43]	RGBD	Stereo	Geometric	1
Harville et al. (2001) [16]	RGBD	Stereo	MoG	1
Kolmogorov et al. (2005) [44]	RGBD	Stereo	MoG	1
Crabb et al. (2008) [45]	RGBD	ToF	Thresholding	2
Guomundsson et al. (2008) [11]	D	ToF	Single Gaussian	1
Wu et al. (2008) [46]	D	IR	Thresholding	1
Frick et al. (2009) [7]	D	ToF	MoG	1
Leens et al. (2009) [47]	RGBD	ToF	ViBe	2
Stormer et al. (2010) [48]	RGBD	ToF	MoG	2
Wang et al. (2010) [49]	RGBD	ToF & Stereo	MoG + Single Gaussian	2
Dondi et al. (2011) [50]	D	ToF	Thresholding	1
Frick et al. (2011) [51]	RGBD	ToF	Thresholding	1
Kawabe et al. (2011) [52]	RGBD	Stereo	MoG	1
Mirante et al. (2011) [53]	RGBD	ToF	Frame diff. + region growing	2
Rougier et al. (2011) [54]	D	Kinect	Single Gaussian	1
Schiller and Koch (2011) [55]	RGBD	ToF	MoG + avg.	2
Stone and Skubic (2011) [56]	D	Kinect	$[d_m, d_M]$	1
Han et al. (2012) [9]	D	Kinect	Frame difference	1
Clapés et al. (2013) [57]	RGBD	Kinect	Single Gaussian	1
Fernandez-Sanchez et al. (2013) [58]	RGBD	Kinect	Codebook	1
Mahbub et al. (2013) [10]	D	Kinect	Frame difference	1
Ottonelli et al. (2013) [59]	RGBD	Stereo	ViBe	2
Zhang et al. (2013) [60]	D	Kinect	Single Gaussian	1
Braham et al. (2014) [61]	D	Kinect	Single Gaussian	2
Camplani and Salgado (2014) [17]	RGBD	Kinect	MoG	2
Camplani et al. (2014) [62]	RGBD	Kinect	MoG	2
Chattopadhyay et al. (2014) [63]	RGBD	Kinect	SOBS	2
Fernandez-Sanchez et al. (2014) [15]	RGBD	Stereo	Codebook	2
Gallego and Pardás (2014) [18]	RGBD	Kinect	MoG	2
Giordano et al. (2014) [64]	RGBD	Kinect	KDE	1
Murgia et al. (2104) [65]	RGBD	Kinect	Codebook	1
Song et al. (2014) [66]	RGBD	Kinect	MoG	2
Boucher et al. (2015) [67]	RGBD	Asus	Mean	1
Cinque et al. (2015) [68]	D	Kinect	Thresholding	1
Huang et al. (2015) [69]	RGBD	Kinect	Thresholding	1
Javed et al. (2015) [70]	RGBD	Stereo	RPCA	2
Nguyen et al. (2015) [71]	RGBD	Kinect	MoG	2
Sun et al. (2105) [72]	RGBD	Kinect	MoG + Single Gaussian	2
Tian et al. (2015) [73]	RGBD	Kinect	RPCA	1
Huang et al. (2016) [19]	RGBD	Kinect	ViBe	2
Liang et al. (2016) [20]	RGBD	Kinect	MoG	2
Palmero et al. (2016) [74]	D	Kinect	MoG	1
Chacon et al. (2017) [75]	RGBD	Kinect	Fuzzy frame diff.	2
De Gregorio and Giordano (2017) [76]	RGBD	Kinect	WiSARD	2
Javed et al. (2017) [77]	RGBD, D	Kinect	RPCA	1
Maddalena and Petrosino (2017) [78]	RGBD	Kinect	SOBS	2
Minematsu et al.(2017) [79]	RGBD	Kinect	ViBe	2
Moyá-Alcover et al. (2017) [32]	RGBD	Kinect	KDE	1
Trabelsi et al. (2017) [80]	RGBD	Kinect & Stereo	KDE	1
Zhou et al. (2017) [81]	RGBD	Kinect	ViBe	1

Crabb et al. [45] propose a method for background substitution, a regularly used effect in TV and video production. Thresholding of depth data coming from a ToF camera, using a user-defined threshold, is adopted to generate a trimap (consisting of background, foreground, and uncertain pixels). Alpha matting values for uncertain pixels, mainly in the borders of the segmented objects, are needed for a natural looking blending of those objects on a different background. They are obtained by cross-bilateral filtering based on color information.

In [11] by Guomundsson et al., 3D multi-person tracking in smart-rooms is tackled. They adopt a single Gaussian model for the range data from a two-modal camera rig (consisting of a ToF range camera and an additional higher resolution grayscale camera) for background subtraction.

In [46], Wu et al. present an algorithm for bi-layer segmentation of natural videos in real time using a combination of infrared, color, and edge information. A prime application of this system is in telepresence, where there is a need to remove the background and replace it with a new one. For each frame, the IR image is used to pre-segment the color image using a simple thresholding technique. This pre-segmentation is adopted to initialize a pentamap, which is then used by graph cuts algorithm to find the final foreground region.

The depth data provided by a ToF camera is used to generate 3D-TV contents by Frick et al. [7]. The MoG algorithm is applied to the depth data to obtain foreground regions, which are then excluded by median filtering to improve background depth map accuracy.

In [47], Leens et al. propose a multi-camera system that combines color and depth data, obtained with a low-resolution ToF camera, for video segmentation. The algorithm applies the ViBe algorithm independently to the color and the depth data. The obtained foreground masks are then combined with logical operations and post-processed with morphological operations.

MoG is also adopted in the algorithm proposed by Stormer et al. [48], where depth and infrared data captured by a ToF camera are combined to detect foreground objects. Two independent background models are built, and each pixel is classified as background or foreground only if the two models matching conditions agree. Very close or overlapping foreground objects are further separated using a depth gradient-based segmentation.

Wang et al. [49] propose TofCut, an algorithm that combines color and depth cues in a unified probabilistic fusion framework and a novel adaptive weighting scheme to control the influence of these two cues intelligently over time. Bilayer segmentation is formulated as a binary labeling problem, whose optimal solution is obtained by minimizing an energy function. The data term evaluates the likelihood of each pixel belonging to the foreground or the background. The contrast term encodes the assumption that segmentation boundaries tend to align with the edges of high contrast. Color and depth foreground and background pixels are modeled through MoGs and single Gaussians, respectively, and their weighting factors are adaptively adjusted based on the discriminative capabilities of their models. The algorithm is also used in an automatic matting system [82] to automatically generate foreground masks, and consequently trimaps, to guide alpha matting.

Dondi et al. [50] propose a matting method using the intensity map generated by ToF cameras. It first segments the distance map based on the corresponding values of the intensity map and then applies region growing to the filtered distance map, to identify and label pixel clusters. A trimap is obtained by eroding the output to select the foreground, dilating it to select foreground, and selecting as indeterminate the remaining contour pixels. The obtained trimap is fed in input to a matting algorithm that refines the result.

Frick et al. [51] use a thresholding technique to separate the foreground from the background in multiple planes of the video volume, for the generation of 3D-TV contents. A posterior trimap-based refinement using hierarchical graph cuts segmentation is further adopted to reduce the artifacts produced by the depth noise.

Kawabe et al. [52] employ stereo cameras to extract pedestrians. Foreground regions are extracted by MoG-based background subtraction and shadow detection using the color data. Then the moving objects are extracted by thresholding the histogram of depth data, computed by stereo matching.

Mirante et al. [53] exploit the information captured by a multi-sensor system consisting of a stereo camera pair with a ToF range sensor. Motion, retrieved by color and depth frame difference, provides the initial ROI mask. The foreground mask is first extracted by region growing in the depth data, where seeds are obtained by the ROI, then refined based on color edges. Finally, a trimap is generated, where uncertain values are those along the foreground contours, and are classified based on color in the CIELab color space.

Rougier et al. [54] explore the Kinect sensor for the application of detecting falls in the elderly. For people detection, the system adopts a single Gaussian depth background model.

Schiller and Koch [55] propose an approach to video matting that combines color information with the depth provided by ToF cameras. Depth keying is adopted to segment moving objects based on depth information, comparing the current depth image with a depth background image (constructed by averaging several ToF-images). MoG is adopted to segment moving objects based on color information. The two segmentations are weighted using two types of reliability measure for depth measurements: the depth variance and the amplitude image of the ToF-camera. The weighted average of the color and depth segmentations is used as matting alpha value for blending foreground and background, while its thresholding (using a user-defined threshold) is used for evaluating moving object segmentation.

Stone and Skubic [56] use only the depth information provided by a Kinect device to extract the foreground. For each pixel, minimum and maximum depth values d_m and d_M are computed by a set of training images to form a background model. For a new frame, each pixel is compared against the background model, and those pixels which lie outside the range $[d_m - 1, d_M + 1]$ are considered foreground.

In [9], Han et al. present a human detection and tracking system for a smart environment application. Background subtraction is applied only on the depth images as frame-by-frame difference, assisted by a clustering algorithm that checks the depth continuity of pixels in the neighborhood of foreground pixels. Once the object has been located in the image, visual features are extracted from the RGB image and are then used for tracking the object in successive frames.

In the surveillance system based on the Kinect proposed by Clapés et al. [57], a per pixel background subtraction technique is presented. The authors propose a background model based on a four-dimensional Gaussian distribution (using color and depth features). Then, user and object candidate regions are detected and recognized using robust statistical approaches.

In the gesture recognition system presented by Mahbub et al. [10], the foreground objects are extracted by the depth data using frame difference.

Ottonelli et al. (2013) [59] refine ViBe segmentation of the color data by adding to the achieved foreground mask a compensation factor computed based on the color and depth data obtained by a stereo camera.

In the object detection system presented by Zhang et al. [60], background subtraction is achieved by single Gaussian modeling of the depth information provided by a Kinect sensor.

Fernandez-Sanchez et al. [58] adopt Codebook as background model and consider data captured by Kinect cameras. They analyze two approaches that differ in the depth integration method: the four-dimensional Codebook (CB4D) considers merely depth as a fourth channel of the background model, while the Depth-Extended Codebook (DECB) adds a joint RGBD fusion method directly into the model. They proved that the latter achieves better results than the former. In [15], the authors consider stereo disparity data, besides color. To get the best of color and depth features, they extend the DECB algorithm through a post-processing stage for mask fusion (DECB-LF), based on morphological reconstruction using the output of the color-based algorithm.

Braham et al. [61] adopt two background models for depth data, separating valid values (modeled by a single Gaussian model) and invalid values (holes). The Gaussian mean is updated to the maximum valid value, while the standard deviation follows a quadratic relationship with respect to the depth. This leads to a depth-dependent foreground/background threshold that enables the model to adapt to the non-uniform noise of range images automatically.

In [17], Camplani and Salgado propose an approach, named CL_W, based on a combination of color and depth classifiers (CL_C and CL_D) and the adoption of the MoG model. The combination of classifiers is based on a weighted average that allows to adaptively modifying the support of each classifier in the ensemble by considering foreground detections in the previous frames and the depth and color edges. For each pixel, the support of each classifier to the final segmentation result is obtained by considering the global edge-closeness probability and the classification labels obtained in the previous frame. In [62], the authors improve their method, proposing a method named MoG-RegPRE, that

builds not only pixel-based but also region-based models from depth and color data, and fuses the models in a mixture of experts fashion to improve the final foreground detection performance.

Chattopadhyay et al. [63] adopt RGBD streams for recognizing gait patterns of individuals. To extract RGBD information of moving objects, they adopt the SOBS model for color background subtraction and use the obtained foreground masks to extract the depth information of people silhouettes from the registered depth frames.

In [18], Gallego and Pardás present a foreground segmentation system that combines color and depth information captured by a Kinect camera to perform a complete Bayesian segmentation between foreground and background classes. The system adopts a combination of spatial-color and spatial-depth region-based MoG models for the foreground, as well as two color and depth pixel-wise MoG models for the background, in a Logarithmic Opinion Pool decision framework used to combine the likelihoods of each model correctly. A post-processing step based on a trimap analysis is also proposed to correct the precision errors that the depth sensor introduces in the object contour.

The algorithm proposed by Giordano et al. in [64] explicitly models the scene background and foreground with a KDE approach in a quantized x-y-hue-saturation-depth space. Foreground segmentation is achieved by thresholding the log-likelihood ratio over the background and foreground probabilities.

Murgia et al. [65] propose an extension of the Codebook model. Similarly to CB4D [58], it includes depth as a fourth channel of the background model but also applies colorimetric invariants to modify the color aspect of the input images, to give them the aspect they would have under canonical illuminants.

In [66], Song et al. model grayscale color and depth values based on MoG. The combination of the two models is based on the product of the likelihoods of the two models.

Boucher et al. [67] initially exploit depth information to achieve a coarse segmentation, using middleware of the adopted ASUS Xtion camera. The obtained mask is refined in uncertain areas (mainly object contours) having high background/foreground contrast, locally modeling colors by their mean value.

Cinque et al. [68] adapt to Kinect data a matting method previously proposed for ToF data. It is based on Otsu thresholding of the depth map and region growing for labeling pixel clusters, assembled to create an alpha map. Edge improvement is obtained by logical OR of the current map with those of the previous four frames.

Huang et al. [69] propose a post-processing framework based on an initial segmentation obtained solely by depth data. Two post-processing steps are proposed: a foreground hole detection step and object boundary refining step. For foreground hole detection, they obtain two weak decisions based on the region color cue and the contour contrast cue, adaptively fused according to their corresponding reliability. For object boundary refinement, they apply weighted fusion of motion probability weighted temporal prior, color likelihood, and smoothness constraints. Therefore, besides handling challenges such as color camouflage, illumination variations, and shadows, the method maintains spatial and temporal consistency of the obtained segmentation, a fundamental issue for the telepresence target application.

Javed et al. [70] propose the DEOR-PCA (Depth Extended Online RPCA) method for background subtraction using binocular cameras. It consists of four main stages: disparity estimation, background modeling, integration, and spatiotemporal constraints. Initially, the range information is obtained using disparity estimation algorithms on a set of stereo pairs. Then, OR-PCA is applied to each of color left image and related disparity image to model the background, separately. The integration adds low-rank and sparse components obtained via OR-PCA to recover the background model and foreground mask from each image. The reconstructed sparse matrix is then thresholded to get the binary foreground mask. Finally, spatiotemporal constraints are applied to remove from the foreground mask most of the noise due to depth information.

In [71], Nguyen et al. present a method where, as an initial offline step, noise is removed from depth data based on a noise model. Background subtraction is then solved by combining RGB and depth features, both modeled by MoG. The fundamental idea in their combination strategy is that when depth measurement is reliable, the segmentation is mainly based on depth information; otherwise, RGB is used as an alternative.

Sun et al. [72] propose a MoG model for color information and a single Gaussian model for depth, together with a color-depth consistency check mechanism driving the updating of the two models. However, experimental results aim at evaluating background estimation, rather than background subtraction.

Tian et al. [73] propose a depth-weighted group-wise PCA-based algorithm, named DG-PCA. The background/foreground separation problem is formulated as a weighted $L_{2,1}$-norm PCA problem with depth-based group sparsity being introduced. Dynamic groups are first generated solely based on depth, and then an iterative solution using depth to define the weights in $L_{2,1}$-norm is developed. The method handles moving cameras through global motion compensation.

In [19], Huang et al. present a method where two separate color and depth background models are based on ViBe, and the two resulting foreground masks are fused by weighted average. The result is further adaptively refined, taking into account multi-cue information (color, depth, and edges) and spatiotemporal consistency (in the neighborhood of foreground pixels in the actual and previous frames).

In [20], Liang et al. propose a method to segment foreground objects based on color and depth data independently, using an existing background subtraction method (in the experiments they choose MoG). They focus on refining the inaccurate results through supervised learning. They extract several features from the source color and depth data in the foreground areas. These features are fed to two independent classifiers (in the experiments they choose random forest [83]) to obtain a better foreground detection.

In [74], Palmero et al. propose a baseline algorithm for human body segmentation using color, depth, and thermal information. To reduce the spatial search space in subsequent steps, the preliminary step is background subtraction, achieved in the depth domain using MoG.

In the method proposed by Chacon et al. [75], named SAFBS (Self-Adapting Fuzzy Background Subtraction), background subtraction is based on two background models for color (in the HSV color space) and depth, providing an initial foreground segmentation by frame differencing. A fuzzy algorithm computes the membership value of each pixel to background or foreground, based on color and depth differences, as well as depth similitude, of the current frame and the background. Temporal and spatial smoothing of the membership values is applied to reduce false alarms due to depth flickering and imprecise measurements around object contours, respectively. The classification result is then employed to update the two background models, using automatically computed learning rates.

De Gregorio and Giordano [76] adapt an existing background modeling method using the WiSARD weightless neural network (WNNs) [41] to the domain of RGBD videos. Color and depth video streams are synchronously but separately modeled by WNNs at each pixel, using a set of initial video frames for network training. In the detection phase, classification is interleaved with re-training on current colors whenever pixels are detected as belonging to the background. Finally, the obtained output masks are combined by an OR operator and post-processed by morphological filtering.

Javed et al. [77] investigate the performance of an online RPCA-based method, named SRPCA, for moving object detection using RGBD videos. The algorithm consists of three main stages: (i) detection of dynamic images to create an input dynamic sequence by discarding motionless video frames; (ii) computation of spatiotemporal graph Laplacians; and (iii) application of RPCA to incorporate the preceding two steps for the separation of background and foreground components. In the experiments, the algorithm is tested by using only intensity, only RGB, and RGBD features, leading to the surprising conclusion that best results are achieved using only intensity features.

The algorithm proposed by Maddalena and Petrosino [78], named RGBD-SOBS, is based on two background models for color and depth information, exploiting a self-organizing neural background model previously adopted for RGB videos [84]. The resulting color and depth detection masks are combined, not only to achieve the final results but also to better guide the selective model update procedure.

Minematsu et al. [79] propose an algorithm, named SCAD, based on a simple combination of the appearance (color) and depth information. The depth background is obtained using, for each pixel, its farthest depth value along the whole video, thus resulting in a batch algorithm. The likelihood of the appearance background is computed using texture-based and RGB-based background subtraction. To reduce false positives due to illumination changes, SCAD roughly detects foreground objects by using texture-based background subtraction. Then, it performs RGB-based background subtraction to improve the results of texture-based background subtraction. Finally, foreground masks are obtained using graph cuts to optimize an energy function which combines the two likelihoods of the background.

Moyá-Alcover et al. [32] construct a scene background model using KDE with a three-dimensional Gaussian kernel. One of the dimensions models depth information, while the other two model normalized chromaticity coordinates. Missing depth data are modeled using a probabilistic strategy to distinguish pixels that belong to the background model from those which are due to foreground objects. Pixels that cannot be classified as background or foreground are placed in the undefined class. Two different implementations are obtained depending on whether undefined pixels are considered as background (GSM_{UB}) or foreground (GSM_{UF}), demonstrating their suitability for scenes where actions happen far or close to the sensor, respectively.

Trabelsi et al. [80] propose the RGBD-KDE algorithm, also based on a scene background model using KDE, but using a two-dimensional Gaussian kernel. One of the dimensions models depth information, while the other models the intensity (average of RGB components). To reduce computational complexity, the Fast Gaussian Transform is adapted to the problem.

Zhou et al. [81] construct color and depth models based on ViBe and fuse the results in a weighting mechanism for the model update that relies on depth reliability.

4. Metrics

The usual way of evaluating the performance of background subtraction algorithms for moving object detection in videos is to pixel-wise compare the computed foreground masks with the corresponding ground truth (GT) foreground masks [26,85] and compute suitable metrics. Metrics frequently adopted for evaluating background subtraction methods in RGBD videos are summarized in Table 2. Here, we report their name (column Name), abbreviation (column Acronym), definition (column Computed as), and whether they should be minimized (↓) or maximized (↑) to have more accurate results (column Better if). All these metrics are defined in terms of the total number of True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) pixels in the whole video. Most of the metrics reported in Table 2 are frequently used for evaluating background subtraction methods in RGB videos [26,85]. The exception is $Si_{\partial\Omega}$, specifically adopted to analyze the errors close to the object boundaries $\partial\Omega$, where depth data is usually very imprecise. In [17], $\partial\Omega$ is defined as an image region made of pixels surrounding the ground truth object boundary and having a distance from it of at most 10 pixels.

Where more than one metric is considered, overall metrics to rank the accuracy of the compared methods are also proposed by some authors [17,32], based on the rankings achieved by the methods according to each of the metrics.

Table 2. Metrics frequently adopted for evaluating background subtraction methods in RGBD videos.

Name	Acronym	Computed as	Better if
Similarity (or Jaccard index)	Si	TP/(TP + FP + FN)	↑
Similarity in $\partial\Omega$	$Si_{\partial\Omega}$	TP/(TP + FP + FN) in $\partial\Omega$	↑
Recall	Rec	TP/(TP + FN)	↑
Specificity	Sp	TN/(TN + FP)	↑
False Positive Rate	FPR	FP/(FP + TN)	↓
False Negative Rate	FNR	FN/(TP + FN)	↓
Percentage of Wrong Classifications	PWC	100 × (FP + FN)/(TP + FN + FP + TN)	↓
Precision	Prec	TP/(TP + FP)	↑
F-Measure	F_1	(2 × Prec × Rec)/(Prec + Rec)	↑

5. Datasets

Several RGBD datasets exist for different tasks, including object detection and tracking, object and scene recognition, human activity analysis, 3D-simultaneous localization and mapping (SLAM), and hand gesture recognition (e.g., see surveys in [24,86,87]). However, depending on the application they have been devised for, they can include single RGBD images instead of videos, or they can supply GTs in the form of bounding boxes, 3D geometries, camera trajectories, 6DOF poses, or dense multi-class labels, rather than GT foreground masks.

In Table 3, we summarize some publicly available RGBD datasets suitable for background subtraction that include videos and, eventually, GT foreground masks. Specifically, we report their acronym, website, and reference publication (column Name & Refs.), the source for depth data (column Source), whether or not they also provide GT foreground masks (column GT masks), the number of videos they include (column No. of videos), some RGBD background subtraction methods adopting them for their evaluation (column Adopted by), and the main application they have been devised for (column Main application).

Table 3. Some publicly available RGBD datasets for background subtraction.

Name & Refs.	Source	GT Masks	No. of Videos	Adopted by	Main Application
GSM [32,88]	Kinect	Yes	7	[32,80]	Background subtraction
Kinect [89,90]	Kinect	No	9	[90]	Background subtraction
MICA-FALL [71,91]	Kinect	No	240	[71]	Analysis of human activities
MULTIVISION Kinect [58,92]	Kinect	Yes	4	[19,58,75,80]	Background subtraction
MULTIVISION Stereo [15,93]	Stereo	Yes	4	[15,70,80]	Background subtraction
Princeton Tracking Benchmark [94,95]	Kinect	No	100	[81]	Tracking
RGB-D Object Detection [17,96]	Kinect	Yes	4	[17,20,62,69,71,75,80]	Background subtraction
RGB-D People [97,98]	Kinect	No	3	[62]	People tracking
SBM-RGBD [33,99]	Kinect	Yes	33	[76–79,100–102]	Background subtraction

The GSM dataset [32] includes seven different sequences designed to test some of the main problems in scene modeling when both color and depth information are used: color camouflage, depth camouflage, color shadows, smooth and sudden illumination changes, and bootstrapping. Each sequence is provided with some hand-labeled GT foreground masks. All the sequences are also included in the SBM-RGBD dataset [33] and accompanied by 56 GT foreground masks.

The Kinect dataset [90] contains nine single person sequences, recorded with a Kinect camera, to show depth and color camouflage situations that are prone to errors in color-depth scenarios.

The MICA-FALL dataset [71] contains RGBD videos for the analysis of human activities, mainly fall detection. Two scenarios are considered for capturing activities that happen at the center field of view of one of the four Kinect sensors or at the cross-view of two or more sensors. Besides color and depth data, accelerometer information and the coordinates of 20 skeleton joints are provided for every frame.

The MULTIVISION dataset consists of two different sets of sequences for the objective evaluation of background subtraction algorithms based on depth information as well as color images. The first set (MULTIVISION Stereo [15]) consists of four sequences recorded by stereo cameras, combined with three different disparity estimation algorithms [103–105]. The sequences are devised to test color saturation, color and depth camouflage, color shadows, low lighting, flickering lights, and sudden illumination changes. The second set (MULTIVISION Kinect [58]) consists of four sequences recorded by a Kinect camera, devised to test out of sensor range depth data, color and depth camouflage, flickering lights, and sudden illumination changes. For all the sequences, some frames have been hand-segmented to provide GT foreground masks. The four MULTIVISION Kinect sequences are also included in the SBM-RGBD dataset [33] and accompanied by 294 GT foreground masks.

The Princeton Tracking Benchmark dataset [95] includes 100 videos covering many realistic cases, such as deformable objects, moving camera, different occlusion conditions, and a variety of clutter backgrounds. The GTs are manual annotations in the form of bounding-boxes drawn around the objects on each frame. One of the sequences (namely, sequence bear_front) is also included in the SBM-RGBD dataset [33] and accompanied by 15 GT foreground masks.

The RGB-D Object Detection dataset [17] includes four different sequences of indoor environments, acquired with a Kinect camera, that contain different demanding situations, such as color and depth camouflage or cast shadows. For each sequence, a hand-labeled ground truth is provided to test foreground/background segmentation algorithms. All the sequences, suitably subdivided and reorganized, are also included in the SBM-RGBD dataset [33] and accompanied by more than 1100 GT foreground masks.

The RGB-D People dataset [98] is devoted to evaluating people detection and tracking algorithms for robotics, interactive systems, and intelligent vehicles. It includes more than 3000 RGBD frames acquired in a university hall from three vertically mounted Kinect sensors. The data contains walking and standing persons seen from different orientations and with different levels of occlusions. Regarding the ground truth, all frames are annotated manually to contain bounding boxes in the 2D depth image space and the visibility status of subjects. Unfortunately, the GT foreground masks built and used in [62] are not available.

The SBM-RGBD dataset [33] is a publicly available benchmarking framework specifically designed to evaluate and compare scene background modeling methods for moving object detection on RGBD videos. It involves the most extensive RGBD video dataset ever made for this specific purpose and also includes videos coming from other datasets, namely, GSM [32], MULTIVISION [58], Princeton Tracking Benchmark [95], RGB-D Object Detection dataset [17], and UR Fall Detection Dataset [106,107]. The 33 videos acquired by Kinect cameras span seven categories, selected to include diverse scene background modeling challenges for moving object detection: illumination changes, color and depth camouflage, intermittent motion, out of sensor depth range, color and depth shadows, and bootstrapping. Depth images 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 signaled in a binary Region-of-Interest (ROI) image and are excluded by the evaluation.

Other publicly available RGBD video datasets are worth mentioning, being equipped with pixel-wise GT foreground masks, which are devoted to specific applications. These include the BIWI RGBD-ID dataset [108,109] and the IPG dataset [110,111], targeted to people re-identification, and the

VAP Trimodal People Segmentation Dataset [74,112], that contains videos captured by thermal, depth, and color sensors, devoted to human body segmentation.

6. Comparisons

Due to the public availability of data, GTs, and results obtained by existing background subtraction algorithms handling RGBD data, five of the RGBD datasets described in Section 5 have been adopted by several authors for benchmarking new algorithms for the problem. Here, we summarize and compare the published results.

6.1. Comparisons on the MULTIVISION Kinect Dataset

Performance comparisons on the MULTIVISION Kinect dataset are reported in Table 4. Here, values for the DECB and the CB4D algorithms by Fernandez et al. [58] and the Codebook algorithm using only color (CB) and only depth (CB-D) are those reported in [58]. Values for the RGBD-KDE algorithm by Trabelsi et al. [80] and the KDE algorithm using only color (C-KDE) and only depth (D-KDE) are those reported in [80]. Values for the RSBS (Random Sampling-based Background Subtraction) algorithm by Huang et al. [19] and the SAFBS algorithm by Chacon et al. [75] are those reported by their authors.

Table 4. Performance results of various background subtraction methods on the RGBD videos of the MULTIVISION Kinect dataset. and σ_{F_1} are the mean and the standard deviation over four GT masks for each video. In boldface, the best results for each metric and each sequence.

Video	Method	F_1	σ_{F_1}	Video	Method	F_1	σ_{F_1}
ChairBox	CB	0.847	0.057	Hallway	CB	0.555	0.189
	C-KDE	0.881	0.062		C-KDE	0.632	**0.052**
	CB-D	0.854	0.058		CB-D	0.770	0.097
	D-KDE	0.933	0.047		D-KDE	**0.923**	0.072
	DECB	0.914	**0.027**		DECB	0.783	0.187
	CB4D	0.886	0.050		CB4D	0.617	0.190
	RSBS	0.895	-		RSBS	0.843	-
	RGBD-KDE	**0.962**	0.032		RGBD-KDE	0.873	0.061
	SAFBS	0.910	-		SAFBS	0.745	-
Shelves	CB	0.699	0.192	Wall	CB	0.843	0.108
	C-KDE	0.885	**0.012**		C-KDE	0.918	0.054
	CB-D	0.835	0.137		CB-D	0.595	0.414
	D-KDE	0.709	0.110		D-KDE	0.665	0.178
	DECB	0.848	0.128		DECB	**0.938**	0.029
	CB4D	0.711	0.205		CB4D	0.868	0.054
	RSBS	0.900	-		RSBS	-	-
	RGBD-KDE	**0.921**	0.033		RGBD-KDE	0.886	**0.009**
	SAFBS	0.894	-		SAFBS	0.930	-

It can be observed that in general, depth alone (i.e., CB-D and D-KDE) allows achieving results better than color alone (i.e., CB and C-KDE), being insensitive to illumination variations (e.g., in sequences ChairBox and Hallway) and color camouflage (e.g., in sequence Hallway). The exception clearly holds for the case of depth camouflage, as in sequence Wall (see Figure 2). For all the sequences, the combined use of both information allows in general to achieve comparable or better performance.

(**a**) (**b**) (**c**)

Figure 2. Wall video from the MULTIVISION Kinect dataset. (**a**) RGB image; (**b**) Depth image; (**c**) GT.

6.2. Comparisons on the MULTIVISION Stereo Dataset

Performance comparisons on the MULTIVISION Stereo dataset are reported in Table 5. Here, values for the DECB-LF algorithm by Fernandez et al. [15], the DECB and the CB4D algorithms by Fernandez et al. [58], and for the Codebook algorithm using only color (CB) and only depth (CB-D) are those reported in [15]. Values for the RGBD-KDE algorithm by Trabelsi et al. [80] and the KDE algorithm using only color (C-KDE) and only depth (D-KDE) are those reported in [80]. Values for the DEOR-PCA algorithm by Javed et al. [70] are those reported by the same authors.

It can be observed that, for all the videos, the combined use of both color and depth information (i.e., DEOR-PCA, DECB, DECB-LF, and RGBD-KDE methods) allows achieving results better than those obtained by color alone (i.e., CB and C-KDE methods) or depth alone (i.e., CB-D and D-KDE methods). Moreover, the difficulty in estimating and discriminating disparities in case of flickering lights (e.g., in LCDScreen and LabDoor videos) and in case of depth camouflage (e.g., in Crossing video) leads depth alone-based methods to obtain worse results as compared to color alone-based methods. Only for sequence Suitcase (see Figure 3), where the main issue is color camouflage, depth alone allows achieving results better than color alone, due to the high accuracy of the estimated depth information.

(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 3. Suitcase video from the MULTIVISION Stereo dataset. (**a**) RGB image; Disparity estimated using: (**b**) Var [103]; (**c**) Phase [104]; and (**d**) SGBM [105]; (**e**) GT.

Table 5. Performance results of various background subtraction methods on the RGBD videos of the MULTIVISION Stereo dataset, using depth from three disparity estimation algorithms (Var [103], Phase [104], and SGBM [105]) or using only color (RGB). F_1 and σ_{F_1} are the mean and the standard deviation over four GT masks for each video. In boldface, the best results for each metric and each video.

Video	Method	Var		Phase		SGBM		RGB	
		F_1	σ_{F_1}	F_1	σ_{F_1}	F_1	σ_{F_1}	F_1	σ_{F_1}
Suitcase	CB							0.520	0.261
	C-KDE							0.688	0.087
	CB-D	0.728	0.106	0.409	0.118	0.560	0.381		
	D-KDE	0.755	**0.068**	0.701	**0.102**	0.783	**0.080**		
	DEOR-PCA	0.826	-	0.431	-	0.413	-		
	DECB	0.766	0.089	0.499	0.147	0.790	0.105		
	DECB-LF	0.750	0.116	0.724	0.116	0.765	0.143		
	RGBD-KDE	**0.847**	0.084	**0.769**	0.057	**0.865**	0.091		
LCDScreen	CB							0.745	0.132
	C-KDE							0.723	0.074
	CB-D	0.400	0.226	0.535	0.099	0.094	0.115		
	D-KDE	0.666	0.172	0.600	0.199	0.782	0.090		
	DEOR-PCA	0.764	-	0.684	-	0.668	-		
	DECB	0.784	0.084	0.639	**0.032**	0.820	0.061		
	DECB-LF	0.803	0.074	0.691	0.071	0.832	0.075		
	RGBD-KDE	**0.904**	**0.042**	**0.820**	0.064	**0.911**	**0.045**		
Crossing	CB							0.653	0.112
	C-KDE							0.741	0.203
	CB-D	0.278	0.334	0.259	0.294	0.387	0.366		
	D-KDE	0.479	0.280	0.507	0.122	0.538	0.252		
	DEOR-PCA	**0.906**	-	0.620	-	0.416	-		
	DECB	0.780	**0.082**	0.636	0.051	0.804	0.048		
	DECB-LF	0.791	**0.082**	0.765	0.111	0.851	0.038		
	RGBD-KDE	0.807	0.089	**0.821**	**0.042**	**0.872**	**0.011**		
LabDoor	CB							0.601	0.176
	C-KDE							0.527	0.081
	CB-D	0.303	0.310	0.217	0.283	0.201	0.321		
	D-KDE	0.573	**0.144**	0.499	0.297	0.476	0.220		
	DEOR-PCA	**0.780**	-	0.547	-	0.572	-		
	DECB	0.548	0.190	0.658	0.137	0.661	0.156		
	DECB-LF	0.674	0.176	**0.673**	0.140	0.691	**0.145**		
	RGBD-KDE	0.614	0.191	0.552	**0.099**	**0.759**	0.182		

6.3. Comparisons on the RGB-D Object Detection Dataset

Performance comparisons on the RGB-D Object Detection dataset are reported in Table 6. Here, values for the two weak color and depth classifiers (CL_C and CL_D) and the weighted color and depth classifier (CL_W) by Camplani and Salgado [17], the four-dimensional MoG model (MoG4D) by Gordon et al. [42], the combined RGB and depth ViBe model (ViBeRGB+D) by Leens et al. [47], and the combined RGB and depth MoG model (MoGRGB+D) by Stormer et al. [48] are those reported in [17]. Values for the RGBD-KDE algorithm by Trabelsi et al. [80] and the KDE algorithm using only color (C-KDE) and only depth (D-KDE) are those reported in [80]. Values for the AMDF (Adaptive Multi-Cue Decision Fusion) algorithm by Huang et al. [69], the RFBS (Refinement Framework for Background Subtraction) algorithm by Liang et al. [20], the EC-RGBD algorithm by Nguyen et al. [71], the enhanced classifier (MoG-RegPRE) by Camplani et al. [62], the GSM_{UB} and GSM_{UF} algorithms by Moyá et al. [32], and the SAFBS algorithm by Chacon et al. [75] are those reported by the related authors.

Good performance can be achieved for color camouflage (ColCamSeq) and shadows (ShSeq), as well as for sequence GenSeq (see Figure 4), which combines different issues (color shadows, color and depth camouflage, and noisy depth data). On the other hand, depth camouflage (DCamSeq) seems to be a problem for most of the methods using depth.

Table 6. Performance results of various background subtraction methods on the RGBD videos of the RGB-D Object Detection dataset. In boldface, the best results for each metric and each sequence.

Video	Method	PWC	FNR	FPR	Si	$SI_{\partial\Omega}$
GenSeq	CL_C	2.38	0.1638	0.0063	0.72	0.55
	C-KDE	2.43	0.1208	0.0088	0.81	0.66
	CL_D	2.06	0.0177	0.0209	0.78	0.42
	D-KDE	0.75	0.0098	0.0107	0.91	0.61
	MoG4D	1.93	**0.0063**	0.0209	0.79	0.45
	ViBeRGB+D	12.39	0.0065	0.1385	0.44	0.12
	MoGRGB+D	2.03	0.1701	**0.0016**	0.79	0.61
	CL_W	1.30	0.0149	0.0127	0.83	0.53
	AMDF	0.94	0.0956	0.0041	-	-
	RFBS	**0.52**	0.0386	0.0045	**0.92**	-
	EC-RGBD	-	-	-	0.87	-
	MoG-RegPRE	0.85	0.0128	0.0079	0.88	-
	GSM_{UB}	1.38	0.0104	0.0144	0.83	**0.78**
	GSM_{UF}	1.30	0.0408	0.0130	0.83	**0.78**
	RGBD-KDE	0.85	0.0115	0.0058	0.87	0.65
	SAFBS	0.60	0.0428	0.0050	**0.92**	-
ColCamSeq	CL_C	39.02	0.8227	0.0227	0.22	0.37
	C-KDE	16.54	0.5972	0.0269	0.51	0.40
	CL_D	2.47	0.0258	0.0238	0.91	**0.78**
	D-KDE	2.68	0.0125	0.0090	0.94	0.77
	MoG4D	3.49	0.0038	0.0613	0.91	0.81
	ViBeRGB+D	6.94	**0.0017**	0.1269	0.81	0.74
	MoGRGB+D	38.47	0.8287	**0.0075**	0.22	0.35
	CL_W	3.20	0.0352	0.0292	0.89	0.77
	AMDF	**1.89**	0.0387	0.0299	-	-
	EC-RGBD	-	-	-	**0.95**	-
	GSM_{UB}	2.30	0.0710	0.0321	0.90	0.52
	GSM_{UF}	2.20	0.0294	0.0436	0.92	0.53
	RGBD-KDE	2.72	0.0299	0.0155	0.83	0.65
DCamSeq	CL_C	1.78	0.1560	0.0095	0.67	0.62
	C-KDE	2.27	0.0689	0.0074	0.74	0.59
	CL_D	3.38	0.4849	0.0064	0.40	0.39
	D-KDE	3.47	0.3012	0.0155	0.52	0.54
	MoG4D	2.11	0.1525	0.0131	0.61	**0.61**
	ViBeRGB+D	9.31	**0.0548**	0.0955	0.30	0.60
	MoGRGB+D	3.57	0.6087	**0.0009**	0.32	0.27
	CL_W	2.46	0.3221	0.0066	0.55	0.51
	AMDF	10.78	0.7686	0.0684	-	-
	GSM_{UB}	1.74	0.2045	0.0046	0.64	0.54
	GSM_{UF}	1.65	0.2206	0.0061	0.65	0.55
	RGBD-KDE	**1.62**	0.1000	0.0049	**0.76**	**0.61**
ShSeq	CL_C	5.37	0.1820	0.0323	0.67	0.63
	C-KDE	3.90	0.0758	0.0259	0.74	0.58
	CL_D	0.98	0.0095	0.0098	0.93	0.67
	D-KDE	0.57	0.0044	**0.0041**	0.88	0.70
	MoG4D	3.94	0.0059	0.0450	0.77	0.66
	ViBeRGB+D	7.15	**0.0001**	0.0834	0.66	0.54
	MoGRGB+D	3.43	0.2351	0.0008	0.75	0.58
	CL_W	0.81	0.0160	0.0068	0.94	0.71
	AMDF	1.46	0.0872	0.0047	-	-
	RFBS	0.47	0.0072	0.0045	**0.96**	-
	EC-RGBD	-	-	-	0.91	-
	GSM_{UB}	0.87	0.0098	0.0088	0.93	0.76
	GSM_{UF}	1.66	0.0014	0.0192	0.89	0.65
	RGBD-KDE	**0.52**	0.0122	0.0049	0.93	**0.80**
	SAFBS	0.65	0.0037	0.0070	0.95	-

(a) (b) (c)

Figure 4. GenSeq video from the RGB-D Object Detection dataset. (**a**) RGB image; (**b**) Depth image; (**c**) GT.

6.4. Comparisons on the GSM Dataset

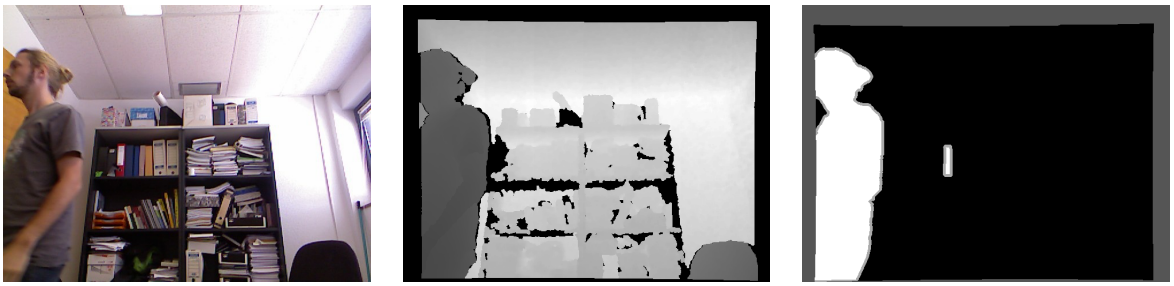
Performance comparisons on the GSM dataset are reported in Table 7. Here, values for the GSM_{UB} and GSM_{UF} algorithms by Moyá et al. [32] are those reported on the dataset website. Values for the RGBD-KDE algorithm by Trabelsi et al. [80] and the KDE algorithm using only color (C-KDE) and only depth (D-KDE) are those reported in [80].

Table 7. Performance results of various background subtraction methods on the RGBD videos of the GSM dataset. In boldface, the best results for each metric and each sequence.

Video	Method	Rec	Sp	FPR	FNR	PWC	F_1	Prec
Sleeping-ds	C-KDE	0.720	0.780	-	-	-	0.705	0.690
	D-KDE	0.790	0.830	-	-	-	0.795	0.800
	GSM_{UB}	0.810	**0.980**	**0.020**	0.190	10.390	0.890	**0.980**
	GSM_{UF}	**0.960**	0.960	0.040	**0.040**	**3.980**	**0.960**	0.950
	RGBD-KDE	0.890	0.880	-	-	-	0.900	0.910
TimeOfDay-ds	C-KDE	0.150	0.480	-	-	-	0.233	0.520
	D-KDE	0.370	0.610	-	-	-	0.469	0.640
	GSM_{UB}	0.000	**1.000**	**0.000**	**0.000**	**0.190**	0.000	0.000
	GSM_{UF}	0.000	**1.000**	**0.000**	**0.000**	0.310	0.000	0.000
	RGBD-KDE	**0.490**	0.750	-	-	-	**0.570**	**0.680**
Cespatx-ds	C-KDE	0.750	0.750	-	-	-	0.755	0.760
	D-KDE	0.910	0.900	-	-	-	0.925	0.940
	GSM_{UB}	0.960	**0.990**	**0.010**	0.040	2.890	0.970	**0.990**
	GSM_{UF}	**0.980**	**0.990**	**0.010**	**0.020**	**1.490**	**0.990**	**0.990**
	RGBD-KDE	0.910	0.940	-	-	-	0.939	0.970
Despatx-ds	C-KDE	0.880	0.910	-	-	-	0.909	0.940
	D-KDE	0.720	0.700	-	-	-	0.739	0.760
	GSM_{UB}	0.940	0.990	0.010	0.060	3.390	0.970	**0.990**
	GSM_{UF}	**0.970**	**1.000**	**0.000**	**0.030**	**0.000**	**0.980**	**0.990**
	RGBD-KDE	0.910	0.940	-	-	-	0.920	0.930
Shadows-ds	C-KDE	0.920	0.920	-	-	-	0.935	0.950
	D-KDE	0.980	0.990	-	-	-	0.985	0.990
	GSM_{UB}	0.960	**1.000**	**0.000**	0.040	1.810	0.980	**1.000**
	GSM_{UF}	0.980	**1.000**	**0.000**	**0.020**	**1.040**	0.990	0.990
	RGBD-KDE	**1.000**	0.990	-	-	-	**1.000**	**1.000**
LightSwitch-ds	GSM_{UB}	**0.000**	**1.000**	**0.000**	**0.000**	**0.110**	**0.000**	**0.000**
	GSM_{UF}	**0.000**	**1.000**	**0.000**	**0.000**	0.340	**0.000**	**0.000**
Bootstraping-ds	C-KDE	0.840	0.880	-	-	-	0.860	0.880
	D-KDE	0.870	0.940	-	-	-	0.908	0.950
	GSM_{UB}	0.740	**1.000**	**0.000**	0.260	6.940	0.850	0.980
	GSM_{UF}	0.850	0.990	0.010	**0.150**	**3.910**	0.910	0.980
	RGBD-KDE	**0.910**	0.980	-	-	-	**0.953**	**1.000**

It can be observed that the compared methods based on the combination of color and depth information robustly deal with all the issues related to RGBD data: intermittent object

motion (Sleeping-ds), illumination changes (TimeOfDay-ds and LightSwitch-ds), color camouflage (Cespatx-ds), depth camouflage (Despatx-ds, see Figure 5), color and depth shadows (Shadows-ds), and bootstrapping (Bootstraping-ds). It should be pointed out that, in the case of TimeOfDay-ds and Ls-ds sequences, the performance analysis should be based on Specificity, FPR, FNR, and PWC, rather than on the other three metrics. Indeed, there are no foreground objects throughout the whole sequences, their rationale being the willingness of not detecting false positives under varying illumination conditions. This leads to having no positive cases in the ground truths and, consequently, to undefined values of Precision, Recall, and F-measure. While for GSM_{UB} and GSM_{UF} values in these undefined cases are set to zero, a different handling must have been adopted for the other compared methods.



(**a**) (**b**) (**c**)

Figure 5. Despatx-ds video from the GSM dataset. (**a**) RGB image; (**b**) Depth image; (**c**) GT.

6.5. Comparisons on the SBM-RGBD Dataset

Performance comparisons on the SBM-RGBD dataset are reported in Tables 8 and 9. Here, values for the RGBD-SOBS and RGB-SOBS algorithms by Maddalena and Petrosino [78], the SRPCA algorithm by Javed et al. [77], the AvgM-D algorithm by Li and Wang [113], the Kim algorithm by Younghee Kim [114], the SCAD algorithm by Minematsu et al. [79], the cwisardH+ algorithm by De Gregorio and Giordano [76], and the MFCN algorithm by Zeng et al. [102], are those reported by the related authors. All the performance measures have been computed using the complete set of GTs and are available at [115].

It can be observed that the deep learning-based MFCN algorithm almost always achieves the best results in all the video categories, in terms of all the metrics. This is certainly possible thanks to the availability of such a wide dataset to train the network. Several conclusions can be drawn for each of the considered challenges by observing the remaining results. Bootstrapping can be a problem when using only color information, especially for selective background subtraction methods (e.g., RGB-SOBS), 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 [116,117] fail and only alternative techniques (e.g., inpainting) can be adopted to recover missing data [118]. Nonetheless, depth information seems to be beneficial for affording the challenge, as reported in Table 8, where accurate results are achieved by most of the methods that exploit depth information.

Table 8. Average results of various background subtraction methods for each category of the SBM-RGBD dataset (Part 1). In boldface, the best results for each metric and each category.

Method	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
MFCN	**0.9866**	**0.9985**	**0.0015**	**0.0134**	**0.2286**	**0.9885**	**0.9876**
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
MFCN	0.9859	**0.9977**	**0.0023**	0.0141	**0.4272**	**0.9893**	**0.9876**
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
MFCN	**0.9870**	**0.9986**	**0.0014**	**0.0130**	**0.2134**	**0.9741**	**0.9804**

As expected, all the methods that exploit depth information achieve high accuracy in case of color camouflage and illumination changes. In the latter case, it should be pointed out that, since this video category includes the two TimeOfDay-ds and Ls-ds sequences of the GSM dataset (without any foreground object), the performance analysis should be based on Specificity, FPR, FNR, and PWC, rather than on the other three metrics (see Section 6.4).

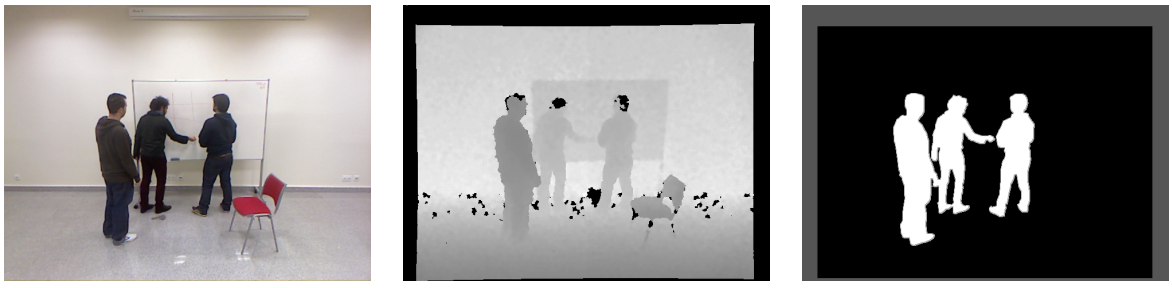
Depth can be beneficial also for detecting and properly handling cases of intermittent motion. Indeed, foreground objects can be easily identified based on their depth, that is lower than that of the background, even when they remain stationary for long time periods. Methods that explicitly exploit this characteristic 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 when only color information is taken into account.

Depth camouflage and out of range (see Figure 6) are among the most challenging issues, at least when information on color is disregarded or not properly combined with depth. Indeed, even though the accuracy of most of the methods is moderately high, several false negatives are produced.

Table 9. Average results of various background subtraction methods for each category of the SBM-RGBD dataset (Part 2). In boldface, the best results for each metric and each category.

Method	Rec	Sp	FPR	FNR	PWC	Prec	F_1
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
MFCN	**0.4986**	**0.9987**	**0.0013**	**0.0014**	**0.1255**	**0.4912**	**0.4949**
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
MFCN	**0.9906**	**0.9987**	**0.0013**	**0.0094**	**0.2466**	**0.9836**	**0.9870**
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
MFCN	**0.9917**	**0.9982**	**0.0018**	**0.0083**	**0.2018**	**0.9613**	**0.9763**
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
MFCN	**0.9893**	**0.9983**	**0.0017**	**0.0107**	**0.2178**	**0.9842**	**0.9867**



(**a**) (**b**) (**c**)

Figure 6. MultiPeople2 video from the SBM-RGBD dataset (OutOfRange category). (**a**) RGB image; (**b**) Depth image; (**c**) GT.

6.6. Summary of the Findings and Open Issues

From the reported comparisons, it can be argued that, generally, most of the issues related to RGB data may be solved by accurate depth information, being insensitive to scene color and illumination conditions (color camouflage, illumination changes, and color shadows) and providing geometric information of the scene (bootstrapping and intermittent motion). This does not hold in cases where

depth measurements or estimation are not sufficiently accurate. However, the combined use of both color and depth information was shown to allow achieving results better than those obtained by color alone or depth alone. Indeed, a clever combination of this information enables the exploitation of depth benefits, at the same time overcoming the issues arising from eventual depth inaccuracies, by exploiting the complimentary color information.

Open issues remain when depth and color information fail to be complimentary. As an example, it has been shown that an object moving on a wall can be detected based on its color, rather than its camouflaged depth. However, what if the object has the same color of the wall? Future research directions should certainly investigate these cases.

7. Conclusions and Future Research Directions

The paper provides a comprehensive review of methods which exploit RGBD data for moving object detection based on background subtraction, a building block for many computer vision applications. The main issues and the existing literature are briefly reviewed. Moreover, the metrics commonly used for the evaluation of these methods and the datasets that are publicly available are summarized. Finally, the most extensive comparison of the existing methods on some datasets is provided, which can serve as a reference for future methods aiming at overcoming the highlighted open issues.

Author Contributions: The authors contributed equally to this work.

Acknowledgments: L.M. acknowledges the INdAM Research group GNCS and the INTEROMICS Flagship Project funded by MIUR, Italy. A.P. acknowledges Project PLI 4.0 Horizon 2020-PON 2014-2020.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bouwmans, T. Traditional and recent approaches in background modeling for foreground detection: An overview. *Comput. Sci. Rev.* **2014**, *11*, 31–66. [CrossRef]
2. Cuevas, C.; Martínez, R.; García, N. Detection of stationary foreground objects: A survey. *Comput. Vis. Image Underst.* **2016**, *152*, 41–57 [CrossRef]
3. Shah, M.; Deng, J.D.; Woodford, B.J. Video background modeling: Recent approaches, issues and our proposed techniques. *Mach. Vis. Appl.* **2014**, *25*, 1105–1119. [CrossRef]
4. Vaswani, N.; Bouwmans, T.; Javed, S.; Narayanamurthy, P. Robust PCA and Robust Subspace Tracking. *arXiv* **2017**, arXiv:1711.09492 [CrossRef]
5. Xu, Y.; Dong, J.; Zhang, B.; Xu, D. Background modeling methods in video analysis: A review and comparative evaluation. *CAAI Trans. Intell. Technol.* **2016**, *1*, 43–60. [CrossRef]
6. Eveland, C.; Konolige, K.; Bolles, R.C. Background modeling for segmentation of video-rate stereo sequences. In Proceedings of the 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA, USA, 25 June 1998; pp. 266–271.
7. Frick, A.; Kellner, F.; Bartczak, B.; Koch, R. Generation of 3D-TV LDV-content with Time-Of-Flight Camera. In Proceedings of the 2009 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, Potsdam, Germany, 4–6 May 2009; pp. 1–4, doi:10.1109/3DTV.2009.5069624. [CrossRef]
8. Greff, K.; Brandão, A.; Krauß, S.; Stricker, D.; Clua, E. A Comparison between Background Subtraction Algorithms using a Consumer Depth Camera. In Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP 2012), Rome, Italy, 24–26 February 2012; Volume 1, pp. 431–436.
9. Han, J.; Pauwels, E.J.; de Zeeuw, P.M.; de With, P.H.N. Employing a RGB-D sensor for real-time tracking of humans across multiple re-entries in a smart environment. *IEEE Trans. Consum. Electron.* **2012**, *58*, 255–263, doi:10.1109/TCE.2012.6227420. [CrossRef]
10. Mahbub, U.; Imtiaz, H.; Roy, T.; Rahman, M.S.; Ahad, M.A.R. A template matching approach of one-shot-learning gesture recognition. *Pattern Recognit. Lett.* **2013**, *34*, 1780–1788. [CrossRef]

11. Guomundsson, S.A.; Larsen, R.; Aanaes, H.; Pardas, M.; Casas, J.R. TOF imaging in Smart room environments towards improved people tracking. In Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Anchorage, AK, USA, 23–28 June 2008; pp. 1–6.
12. Xia, L.; Chen, C.C.; Aggarwal, J.K. Human detection using depth information by Kinect. In Proceedings of the 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2011), Colorado Springs, CO, USA, 20–25 June 2011; pp. 15–22, doi:10.1109/CVPRW.2011.5981811. [CrossRef]
13. Almazan, E.J.; Jones, G.A. Tracking People across Multiple Non-overlapping RGB-D Sensors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 831–837.
14. Galanakis, G.; Zabulis, X.; Koutlemanis, P.; Paparoulis, S.; Kouroumalis, V. Tracking Persons Using a Network of RGBD Cameras. In Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '14), Rhodes, Greece, 27–30 May 2014; ACM: New York, NY, USA, 2014; p. 63.
15. Fernandez-Sanchez, E.J.; Rubio, L.; Diaz, J.; Ros, E. Background subtraction model based on color and depth cues. *Mach. Vis. Appl.* **2014**, *25*, 1211–1225. [CrossRef]
16. Harville, M.; Gordon, G.; Woodfill, J. Foreground segmentation using adaptive mixture models in color and depth. In Proceedings of the IEEE Workshop on Detection and Recognition of Events in Video, Vancouver, BC, Canada, 8 July 2001; pp. 3–11, doi:10.1109/EVENT.2001.938860. [CrossRef]
17. Camplani, M.; Salgado, L. Background foreground segmentation with RGB-D Kinect data: An efficient combination of classifiers. *J. Vis. Commun. Image Represent.* **2014**, *25*, 122–136. [CrossRef]
18. Gallego, J.; Pardás, M. Region based foreground segmentation combining color and depth sensors via logarithmic opinion pool decision. *J. Vis. Commun. Image Represent.* **2014**, *25*, 184–194. [CrossRef]
19. Huang, J.; Wu, H.; Gong, Y.; Gao, D. Random sampling-based background subtraction with adaptive multi-cue fusion in RGBD videos. In Proceedings of the 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Datong, China, 15–17 October 2016; pp. 30–35, doi:10.1109/CISP-BMEI.2016.7852677. [CrossRef]
20. Liang, Z.; Liu, X.; Liu, H.; Chen, W. A refinement framework for background subtraction based on color and depth data. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 271–275, doi:10.1109/ICIP.2016.7532361. [CrossRef]
21. Cruz, L.; Lucio, D.; Velho, L. Kinect and RGBD Images: Challenges and Applications. In Proceedings of the 2012 25th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials, Ouro Preto, Brazil, 22–25 August 2012; pp. 36–49.
22. Zhang, Z. Microsoft Kinect Sensor and Its Effect. *IEEE MultiMedia* **2012**, *19*, 4–10. [CrossRef]
23. Han, J.; Shao, L.; Xu, D.; Shotton, J. Enhanced Computer Vision With Microsoft Kinect Sensor: A Review. *IEEE Trans. Cybern.* **2013**, *43*, 1318–1334, doi:10.1109/TCYB.2013.2265378. [CrossRef] [PubMed]
24. Camplani, M.; Paiement, A.; Mirmehdi, M.; Damen, D.; Hannuna, S.; Burghardt, T.; Tao, L. Multiple human tracking in RGB-depth data: A survey. *IET Comput. Vis.* **2017**, *11*, 265–285. [CrossRef]
25. Toyama, K.; Krumm, J.; Brumitt, B.; Meyers, B. Wallflower: Principles and practice of background maintenance. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 1, pp. 255–261, doi:10.1109/ICCV.1999.791228. [CrossRef]
26. Goyette, N.; Jodoin, P.M.; Porikli, F.; Konrad, J.; Ishwar, P. A novel video dataset for change detection benchmarking. *IEEE Trans. Image Process.* **2014**, *23*, 4663–4679. [CrossRef] [PubMed]
27. Zanuttigh, P.; Marin, G.; Dal Mutto, C.; Dominio, F.; Minto, L.; Cortelazzo, G.M. *Time-of-Flight and Structured Light Depth Cameras*; Springer International Publishing: Berlin/Heidelberg, Germany, 2016.
28. Hu, X.; Mordohai, P. A Quantitative Evaluation of Confidence Measures for Stereo Vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2121–2133. [PubMed]
29. Kolb, A.; Barth, E.; Koch, R.; Larsen, R. Time-of-Flight Cameras in Computer Graphics. *Comput. Graph. Forum* **2010**, *29*, 141–159. [CrossRef]
30. Daneshmand, M.; Helmi, A.; Avots, E.; Noroozi, F.; Alisinanoglu, F.; Arslan, H.S.; Gorbova, J.; Haamer, R.E.; Ozcinar, C.; Anbarjafari, G. 3D Scanning: A Comprehensive Survey. *arXiv* **2018**, arXiv:1801.08863. [CrossRef]

31. Khoshelham, K.; Elberink, S.O. Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications. *Sensors* **2012**, *12*, 1437–1454. [CrossRef] [PubMed]
32. Moyà-Alcover, G.; Elgammal, A.; i Capó, A.J.; Varona, J. Modeling depth for nonparametric foreground segmentation using RGBD devices. *Pattern Recognit. Lett.* **2017**, *96*, 76–85. [CrossRef]
33. Camplani, M.; Maddalena, L.; Moyá Alcover, G.; Petrosino, A.; Salgado, L. A Benchmarking Framework for Background Subtraction in RGBD Videos. In Proceedings of the New Trends in Image Analysis and Processing (ICIAP 2017), Catania, Italy, 11–15 September 2017; Battiato, S., Farinella, G.M., Leo, M., Gallo, G., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; pp. 219–229.
34. Kim, K.; Chalidabhongse, T.H.; Harwood, D.; Davis, L. Real-time foreground-background segmentation using codebook model. *Real-Time Imaging* **2005**, *11*, 172–185, doi:10.1016/j.rti.2004.12.004. [CrossRef]
35. Elgammal, A.M.; Harwood, D.; Davis, L.S. Non-parametric Model for Background Subtraction. In Proceedings of the 6th European Conference on Computer Vision, Dublin, Ireland, 26 June–1 July 2000; Springer: Berlin/Heidelberg, Germany, 2000; pp. 751–767.
36. Stauffer, C.; Grimson, W.E.L. Adaptive background mixture models for real-time tracking. In Proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Fort Collins, CO, USA, 23–25 June 1999; Volume 2, p. 252, doi:10.1109/CVPR.1999.784637. [CrossRef]
37. Candès, E.J.; Li, X.; Ma, Y.; Wright, J. Robust Principal Component Analysis? *J. ACM* **2011**, *58*, 11. [CrossRef]
38. Maddalena, L.; Petrosino, A. A Self-Organizing Approach to Background Subtraction for Visual Surveillance Applications. *IEEE Trans. Image Process.* **2008**, *17*, 1168–1177. [CrossRef] [PubMed]
39. Wren, C.R.; Azarbayejani, A.; Darrell, T.; Pentland, A.P. Pfinder: Real-time tracking of the human body. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 780–785. [CrossRef]
40. Barnich, O.; Droogenbroeck, M.V. ViBE: A powerful random technique to estimate the background in video sequences. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 945–948, doi:10.1109/ICASSP.2009.4959741. [CrossRef]
41. Aleksander, I.; Thomas, W.; Bowden, P. WISARD·a radical step forward in image recognition. *Sens. Rev.* **1984**, *4*, 120–124. [CrossRef]
42. Gordon, G.G.; Darrell, T.; Harville, M.; Woodfill, J. Background Estimation and Removal Based on Range and Color. In Proceedings of the 1999 Conference on Computer Vision and Pattern Recognition (CVPR '99), Ft. Collins, CO, USA, 23–25 June 1999; pp. 2459–2464, doi:10.1109/CVPR.1999.784721. [CrossRef]
43. Ivanov, Y.; Bobick, A.; Liu, J. Fast Lighting Independent Background Subtraction. *Int. J. Comput. Vis.* **2000**, *37*, 199–207, doi:10.1023/A:1008107805263. [CrossRef]
44. Kolmogorov, V.; Criminisi, A.; Blake, A.; Cross, G.; Rother, C. Bi-layer segmentation of binocular stereo video. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 2, p. 1186, doi:10.1109/CVPR.2005.90. [CrossRef]
45. Crabb, R.; Tracey, C.; Puranik, A.; Davis, J. Real-time foreground segmentation via range and color imaging. In Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Anchorage, AK, USA, 23–28 June 2008; pp. 1–5, doi:10.1109/CVPRW.2008.4563170. [CrossRef]
46. Wu, Q.; Boulanger, P.; Bischof, W.F. Robust Real-Time Bi-Layer Video Segmentation Using Infrared Video. In Proceedings of the 2008 Canadian Conference on Computer and Robot Vision, Windsor, ON, Canada, 28–30 May 2008; pp. 87–94, doi:10.1109/CRV.2008.7. [CrossRef]
47. Leens, J.; Piérard, S.; Barnich, O.; Van Droogenbroeck, M.; Wagner, J.M. Combining Color, Depth, and Motion for Video Segmentation. In Proceedings of the Computer Vision Systems: 7th International Conference on Computer Vision Systems (ICVS 2009), Liège, Belgium, 13–15 October 2009; Fritz, M., Schiele, B., Piater, J.H., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 104–113, doi:10.1007/978-3-642-04667-4_11. [CrossRef]
48. Stormer, A.; Hofmann, M.; Rigoll, G. Depth gradient based segmentation of overlapping foreground objects in range images. In Proceedings of the 2010 13th International Conference on Information Fusion, Edinburgh, UK, 26–29 July 2010; pp. 1–4, doi:10.1109/ICIF.2010.5712108. [CrossRef]
49. Wang, L.; Zhang, C.; Yang, R.; Zhang, C. TofCut: Towards Robust Real-time Foreground Extraction using Time-of-flight Camera. In Proceedings of the 3DPVT, Paris, France, 17–20 May 2010.

50. Dondi, P.; Lombardi, L. Fast Real-time Segmentation and Tracking of Multiple Subjects by Time-of-Flight Camera—A New Approach for Real-time Multimedia Applications with 3D Camera Sensor. In Proceedings of the Sixth International Conference on Computer Vision Theory and Applications (VISAPP 2011), Vilamoura, Portugal, 5–7 March 2011; pp. 582–587.
51. Frick, A.; Franke, M.; Koch, R. Time-Consistent Foreground Segmentation of Dynamic Content from Color and Depth Video. In Proceedings of the Pattern Recognition: 33rd DAGM Symposium, Frankfurt/Main, Germany, 31 August–2 September 2011; Mester, R., Felsberg, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 296–305, doi:10.1007/978-3-642-23123-0_30. [CrossRef]
52. Kawabe, M.; Tan, J.K.; Kim, H.; Ishikawa, S.; Morie, T. Extraction of individual pedestrians employing stereo camera images. In Proceedings of the 2011 11th International Conference on Control, Automation and Systems, Gyeonggi-do, Korea, 26–29 October 2011; pp. 1744–1747.
53. Mirante, E.; Georgiev, M.; Gotchev, A. A fast image segmentation algorithm using color and depth map. In Proceedings of the 2011 3DTV Conference: The True Vision—Capture, Transmission and Display of 3D Video (3DTV-CON), Antalya, Turkey, 16–18 May 2011; pp. 1–4.
54. Rougier, C.; Auvinet, E.; Rousseau, J.; Mignotte, M.; Meunier, J. Fall Detection from Depth Map Video Sequences. In Proceedings of the Toward Useful Services for Elderly and People with Disabilities: 9th International Conference on Smart Homes and Health Telematics (ICOST 2011), Montreal, QC, Canada, 20–22 June 2011; Abdulrazak, B., Giroux, S., Bouchard, B., Pigot, H., Mokhtari, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 121–128, doi:10.1007/978-3-642-21535-3_16. [CrossRef]
55. Schiller, I.; Koch, R. Improved Video Segmentation by Adaptive Combination of Depth Keying and Mixture-of-Gaussians. In Proceedings of the Image Analysis—17th Scandinavian Conference (SCIA 2011), Ystad, Sweden, 23–25 May 2011; pp. 59–68, doi:10.1007/978-3-642-21227-7_6. [CrossRef]
56. Stone, E.E.; Skubic, M. Evaluation of an inexpensive depth camera for passive in-home fall risk assessment. In Proceedings of the 2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops, Dublin, Ireland, 23–26 May 2011; pp. 71–77, doi:10.4108/icst.pervasivehealth.2011.246034. [CrossRef]
57. Clapés, A.; Reyes, M.; Escalera, S. Multi-modal user identification and object recognition surveillance system. *Pattern Recognit. Lett.* **2013**, *34*, 799–808. [CrossRef]
58. Fernandez-Sanchez, E.J.; Diaz, J.; Ros, E. Background Subtraction Based on Color and Depth Using Active Sensors. *Sensors* **2013**, *13*, 8895–8915. [CrossRef] [PubMed]
59. Ottonelli, S.; Spagnolo, P.; Mazzeo, P.L.; Leo, M. Improved video segmentation with color and depth using a stereo camera. In Proceedings of the 2013 IEEE International Conference on Industrial Technology (ICIT), Cape Town, South Africa, 25–28 February 2013; pp. 1134–1139.
60. Xucong Zhang, X.W.; Jia, Y. The visual internet of things system based on depth camera. In Proceedings of the Chinese Intelligent Automation Conference (CIAC 2013), Yangzhou, China, 23–25 August 2013.
61. Braham, M.; Lejeune, A.; Droogenbroeck, M.V. A physically motivated pixel-based model for background subtraction in 3D images. In Proceedings of the 2014 International Conference on 3D Imaging (IC3D), Liege, Belgium, 9–10 December 2014; pp. 1–8, doi:10.1109/IC3D.2014.7032591. [CrossRef]
62. Camplani, M.; del Blanco, C.R.; Salgado, L.; Jaureguizar, F.; García, N. Multi-sensor background subtraction by fusing multiple region-based probabilistic classifiers. *Pattern Recognit. Lett.* **2014**, *50*, 23–33, doi:10.1016/j.patrec.2013.09.022. [CrossRef]
63. Chattopadhyay, P.; Roy, A.; Sural, S.; Mukhopadhyay, J. Pose Depth Volume extraction from RGB-D streams for frontal gait recognition. *J. Vis. Commun. Image Represent.* **2014**, *25*, 53 – 63. [CrossRef]
64. Giordano, D.; Palazzo, S.; Spampinato, C. Kernel Density Estimation Using Joint Spatial-Color-Depth Data for Background Modeling. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 4388–4393, doi:10.1109/ICPR.2014.751. [CrossRef]
65. Murgia, J.; Meurie, C.; Ruichek, Y. An Improved Colorimetric Invariants and RGB-Depth-Based Codebook Model for Background Subtraction Using Kinect. In Proceedings of the Human-Inspired Computing and Its Applications, Tuxtla Gutiérrez, Mexico, 16–22 November 2014; Gelbukh, A., Espinoza, F.C., Galicia-Haro, S.N., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 380–392.

66. Song, Y.M.; Noh, S.; Yu, J.; Park, C.W.; Lee, B.G. Background subtraction based on Gaussian mixture models using color and depth information. In Proceedings of the 2014 International Conference on Control, Automation and Information Sciences (ICCAIS 2014), Gwangju, South Korea, 2–5 December 2014; pp. 132–135, doi:10.1109/ICCAIS.2014.7020544. [CrossRef]
67. Boucher, A.; Martinot, O.; Vincent, N. Depth Camera to Improve Segmenting People in Indoor Environments—Real Time RGB-Depth Video Segmentation. In Proceedings of the 10th International Conference on Computer Vision Theory and Applications, Berlin, Germany, 11–14 March 2015; Volume 3, pp. 55–62.
68. Cinque, L.; Danani, A.; Dondi, P.; Lombardi, L. Real-Time Foreground Segmentation with Kinect Sensor. In Proceedings of the Image Analysis and Processing (ICIAP 2015), Genoa, Italy, 11–17 September 2015; Murino, V., Puppo, E., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 56–65.
69. Huang, M.; Chen, Y.; Ji, W.; Miao, C. Accurate and Robust Moving-Object Segmentation for Telepresence Systems. *ACM Trans. Intell. Syst. Technol.* **2015**, *6*, 17. [CrossRef]
70. Javed, S.; Bouwmans, T.; Jung, S.K. Depth extended online RPCA with spatiotemporal constraints for robust background subtraction. In Proceedings of the Korea-Japan Workshop on Frontiers of Computer Vision (FCV 2015), Mokpo, South Korea, 28–30 January 2015; pp. 1–6.
71. Nguyen, V.T.; Vu, H.; Tran, T.H. An Efficient Combination of RGB and Depth for Background Subtraction. In Proceedings of the Some Current Advanced Researches on Information and Computer Science in Vietnam: Post-proceedings of The First NAFOSTED Conference on Information and Computer Science, Ha Noi, Vietnam, 13–14 March 2014; Dang, Q.A., Nguyen, X.H., Le, H.B., Nguyen, V.H., Bao, V.N.Q., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 49–63, doi:10.1007/978-3-319-14633-1_4. [CrossRef]
72. Sun, B.; Tillo, T.; Xu, M. Adaptive Model for Background Extraction Using Depth Map. In Proceedings of the Advances in Multimedia Information Processing (PCM 2015): 16th Pacific-Rim Conference on Multimedia, Gwangju, South Korea, 16–18 September 2015; Ho, Y.S., Sang, J., Ro, Y.M., Kim, J., Wu, F., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; Part II, pp. 419–427, doi:10.1007/978-3-319-24078-7_42. [CrossRef]
73. Tian, D.; Mansour, H.; Vetro, A. Depth-weighted group-wise principal component analysis for video foreground/background separation. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 Septmber 2015; pp. 3230–3234, doi:10.1109/ICIP.2015.7351400. [CrossRef]
74. Palmero, C.; Clapés, A.; Bahnsen, C.; Møgelmose, A.; Moeslund, T.B.; Escalera, S. Multi-modal RGB–Depth–Thermal Human Body Segmentation. *Int. J. Comput. Vis.* **2016**, *118*, 217–239. [CrossRef]
75. Chacon-Murguia, M.I.; Orozco-Rodríguez, H.E.; Ramirez-Quintana, J.A. Self-Adapting Fuzzy Model for Dynamic Object Detection Using RGB-D Information. *IEEE Sens. J.* **2017**, *17*, 7961–7970. [CrossRef]
76. De Gregorio, M.; Giordano, M. WiSARD-based learning and classification of background in RGBD videos. In Proceedings of the New Trends in Image Analysis and Processing (ICIAP 2017), Catania, Italy, 11–15 September 2017; Battiato, S., Farinella, G.M., Leo, M., Gallo, G., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2017.
77. Javed, S.; Bouwmans, T.; Sultana, M.; Jung, S.K. Moving Object Detection on RGB-D Videos Using Graph Regularized Spatiotemporal RPCA. In Proceedings of the New Trends in Image Analysis and Processing (ICIAP 2017), Catania, Italy, 11–15 September 2017; Battiato, S., Farinella, G.M., Leo, M., Gallo, G., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; pp. 230–241.
78. Maddalena, L.; Petrosino, A. Exploiting Color and Depth for Background Subtraction. In Proceedings of the New Trends in Image Analysis and Processing (ICIAP 2017), Catania, Italy, 11–15 September 2017; Battiato, S., Farinella, G.M., Leo, M., Gallo, G., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; pp. 254–265.
79. Minematsu, T.; Shimada, A.; Uchiyama, H.; Taniguchi, R. Simple Combination of Appearance and Depth for Foreground Segmentation. In Proceedings of the New Trends in Image Analysis and Processing (ICIAP 2017), Catania, Italy, 11–15 September 2017; Battiato, S., Farinella, G.M., Leo, M., Gallo, G., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2017.
80. Trabelsi, R.; Jabri, I.; Smach, F.; Bouallegue, A. Efficient and fast multi-modal foreground-background segmentation using RGBD data. *Pattern Recognit. Lett.* **2017**, *97*, 13–20. [CrossRef]

81. Zhou, X.; Liu, X.; Jiang, A.; Yan, B.; Yang, C. Improving Video Segmentation by Fusing Depth Cues and the Visual Background Extractor (ViBe) Algorithm. *Sensors* **2017**, *17*, 1177. [CrossRef] [PubMed]
82. Wang, L.; Gong, M.; Zhang, C.; Yang, R.; Zhang, C.; Yang, Y.H. Automatic Real-Time Video Matting Using Time-of-Flight Camera and Multichannel Poisson Equations. *Int. J. Comput. Vis.* **2012**, *97*, 104–121. [CrossRef]
83. Dollár, P.; Zitnick, C.L. Fast Edge Detection Using Structured Forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1558–1570. [CrossRef] [PubMed]
84. Maddalena, L.; Petrosino, A. The SOBS algorithm: What are the limits? In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Providence, RI, USA, 16–21 June 2012; pp. 21–26, doi:10.1109/CVPRW.2012.6238922.
85. Goyette, N.; Jodoin, P.M.; Porikli, F.; Konrad, J.; Ishwar, P. Changedetection.net: A new change detection benchmark dataset. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Providence, RI, USA, 16–21 June 2012; pp. 1–8, doi:10.1109/CVPRW.2012.6238919. [CrossRef]
86. Firman, M. RGBD Datasets: Past, Present and Future. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA , 26 June–1 July 2016; pp. 661–673.
87. Cai, Z.; Han, J.; Liu, L.; Shao, L. RGB-D datasets using microsoft kinect or similar sensors: A survey. *Multimedia Tools Appl.* **2017**, *76*, 4313–4355. [CrossRef]
88. GSM Dataset. Available online: http://gsm.uib.es/ (accessed on 15 May 2018).
89. Kinect Database. Available online: https://imatge.upc.edu/web/resources/kinect-database-foreground-segmentation (accessed on 15 May 2018).
90. Gallego, J. Parametric Region-Based Foreground Segmentation in Planar and Multi-View Sequences. Ph.D. Thesis, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, 2013.
91. MICA-FALL Dataset. Available online: http://mica.edu.vn/perso/Tran-Thi-Thanh-Hai/MFD.html (accessed on 15 May 2018).
92. MULTIVISION Kinect Dataset. Available online: http://atcproyectos.ugr.es/mvision/index.php?option=com_content&view=article&id=45&Itemid=57 (accessed on 15 May 2018).
93. MULTIVISION Stereo Dataset. Available online: http://atcproyectos.ugr.es/mvision/index.php?option=com_content&view=article&id=45&Itemid=57 (accessed on 15 May 2018).
94. Princeton Tracking Benchmark Dataset. Available online: http://tracking.cs.princeton.edu/dataset.html (accessed on 15 May 2018).
95. Song, S.; Xiao, J. Tracking Revisited Using RGBD Camera: Unified Benchmark and Baselines. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 233–240.
96. RGB-D Object Detection Dataset. Available online: http://eis.bristol.ac.uk/~mc13306/ (accessed on 15 May 2018).
97. RGB-D People Dataset. Available online: http://www2.informatik.uni-freiburg.de/~spinello/RGBD-dataset.html (accessed on 15 May 2018).
98. Spinello, L.; Arras, K.O. People detection in RGB-D data. In Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, CA, USA, 25–30 September 2011; pp. 3838–3843, doi:10.1109/IROS.2011.6095074. [CrossRef]
99. SBM-RGBD Dataset. Available online: http://rgbd2017.na.icar.cnr.it/SBM-RGBDdataset.html(accessed on 15 May 2018).
100. Kim, Y. Kim Method. Unpublished work, 2017.
101. Li, G.L.; Wang, X. AvgM-D. Unpublished work, 2017.
102. Zeng, D.; Zhu, M. Background Subtraction Using Multiscale Fully Convolutional Network. *IEEE Access* **2018**. [CrossRef]
103. Ralli, J.; Díaz, J.; Ros, E. Spatial and temporal constraints in variational correspondence methods. *Mach. Vis. Appl.* **2013**, *24*, 275–287. [CrossRef]
104. Hirschmuller, H. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 328–341. [CrossRef] [PubMed]

105. Tomasi, M.; Vanegas, M.; Barranco, F.; Daz, J.; Ros, E. Massive Parallel-Hardware Architecture for Multiscale Stereo, Optical Flow and Image-Structure Computation. *IEEE Trans. Circuits Syst. Video Technol.* **2012**, *22*, 282–294. [CrossRef]
106. Kwolek, B.; Kepski, M. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Comput. Methods Prog. Biomed.* **2014**, *117*, 489–501. [CrossRef] [PubMed]
107. UR Fall Detection Dataset. Available online: http://fenix.univ.rzeszow.pl/~mkepski/ds/uf.html (accessed on 15 May 2018).
108. BIWI RGBD-ID Dataset. Available online: http://robotics.dei.unipd.it/reid/index.php/8-dataset/2-overview-biwi (accessed on 15 May 2018).
109. Munaro, M.; Fossati, A.; Basso, A.; Menegatti, E.; Van Gool, L. One-Shot Person Re-identification with a Consumer Depth Camera. In *Person Re-Identification*; Gong, S., Cristani, M., Yan, S., Loy, C.C., Eds.; Springer: London, UK, 2014; pp. 161–181.
110. IPG Dataset. Available online: http://www.gpiv.upv.es/kinect_data/ (accessed on 15 May 2018).
111. Albiol, A.; Albiol, A.; Oliver, J.; Mossi, J.M. Who is who at different cameras: people re-identification using depth cameras. *IET Comput. Vis.* **2012**, *6*, 378–387. [CrossRef]
112. VAP Trimodal People Segmentation Dataset. Available online: http://www.vap.aau.dk/ (accessed on 15 May 2018).
113. Li, G.L.; Wang, X. AvgM-D algorithm. Unpublished work, 2017.
114. Kim, Y. Kim Algorithm. Unpublished work, 2017.
115. SBM-RGBD Challenge Results. Available online: http://rgbd2017.na.icar.cnr.it/SBM-RGBDchallengeResults.html (accessed on 15 May 2018).
116. Bouwmans, T.; Maddalena, L.; Petrosino, A. Scene background initialization: A taxonomy. *Pattern Recognit. Lett.* **2017**, *96*, 3–11. [CrossRef]
117. Kajo, I.; Kamel, N.; Ruichek, Y.; Malik, A.S. SVD-Based Tensor-Completion Technique for Background Initialization. *IEEE Trans. Image Process.* **2018**, *27*, 3114–3126. [CrossRef]
118. Maddalena, L.; Petrosino, A. Background Model Initialization for Static Cameras. In *Background Modeling and Foreground Detection for Video Surveillance*; Bouwmans, T., Porikli, F., Hoferlin, B., Vacavant, A., Eds.; Chapman & Hall/CRC: New York, NY, USA, 2014; pp. 3–1–3–16.