M. BRAHAM, A. LEJEUNE and M. VAN DROOGENBROECK. **A physically motivated pixel-based model for background subtraction in 3D images**. In *International Conference on 3D Imaging (IC3D)*, Liège, Belgium, December 2014.

A PHYSICALLY MOTIVATED PIXEL-BASED MODEL FOR BACKGROUND SUBTRACTION IN 3D IMAGES

Marc Braham, Antoine Lejeune and Marc Van Droogenbroeck

INTELSIG Laboratory, Montefiore Institute, University of Liège, Belgium

ABSTRACT

This paper proposes a new pixel-based background subtraction technique, applicable to range images, to detect motion. Our method exploits the physical meaning of depth information, which leads to an improved background/foreground segmentation and the instantaneous suppression of ghosts that would appear on color images.

In particular, our technique considers certain characteristics of depth measurements, such as failures for certain pixels or the non-uniformity of the spatial distribution of noise in range images, to build an improved pixel-based background model. Experiments show that incorporating specificities related to depth measurements allows us to propose a method whose performance is increased with respect to other state-of-the-art methods.

Index Terms— Range camera; background subtraction; motion detection; Kinect; depth imagery; video surveillance

1. INTRODUCTION

Motion detection is one of the most essential tasks in computer vision, especially for dealing with real-time video streams. Video surveillance, event monitoring, people counting and face recognition are some examples of applications using motion detection as a pre-processing step.

One of the most straightforward approaches to motion detection consists in building a model of the static scene, which is named *background*, and comparing it with the values of each pixel of an image, one by one. If, for a given pixel, there is a noticeable difference, then the pixel is supposed to belong to an object in motion and is labeled as being in the *foreground*. Background subtraction is thus a two class segmentation technique (foreground or background).

While the principle is simple, there are many practical problems for color images because of many known issues such as sudden lighting changes, camouflage effects, or the appearance of ghosts due to static objects in the scene that start moving. Cameras that measure depth instead of color therefore offer an interesting alternative for motion detection because depth is not pruned to the same limitations and because the signal itself has a different physical meaning. For example, 3D cameras such as the Kinect [8] are almost insensitive to sudden illumination changes and they are unrelated to the true color of objects (to a large degree). Background models obtained for range images are thus expected to be more stable, as long as they take the physical nature of depth into account.

Despite the inherent advantages of range imaging compared to color imaging, it is ineffective to only transpose the principles of background subtraction methods tailored for color images to depth images as such. Although this approach is favored almost exclusively by researchers, results are below the achievable performance. We found that considering the physical nature of depth and its measurement, such as measurement failures or the relationship between measurement uncertainty and depth, helps to increase the segmentation quality significantly.

This paper proposes a new background subtraction technique tailored for depth imaging; our method was further refined for the Kinect camera. It is organized as follows. Section 2 reviews the state-of-the-art for motion detection in range images. Our method is detailed in Section 3. Section 4 evaluates our method and compares it to other techniques. Finally, Section 5 concludes the paper.

2. RELATED WORK

Most of the literature work for motion detection has been dedicated to color imaging. As emphasized in some reviewing papers [2, 13], the numerous color-based techniques exploit the statistical properties of colors and specific methodologies and datasets were developed to compare and evaluate all these methods for colors. In comparison, motion detection for range images is almost unexplored, as mentioned by Greff *et al.* [9]. In some papers on 3D imaging, authors have mainly developed post-processing techniques and they did not focus on the raw data itself. For example, Schwarz *et al.* [17] use a basic technique that assumes a static background and the absence of foreground objects during the initialization phase. Jansen *et al.* [12] apply the method of the exponential filter to 3D images, while Guðmundsson *et al.* [10] rely on one adaptive Gaussian per pixel. Other authors use the well-known mixture of Gaussians model [20, 23], but they ignore the particular characteristics of depth information.

Since the commercialization of the Kinect camera in 2010, there have been some efforts to combine color and depth information provided by the camera. Several algorithms for RGB-D background subtraction have been proposed. Often, these algorithms merge the informations provided by the two separate RGB and D sensors, and they consist in applying classical methods to depth images. For example, Lorenzo-Navarro *et al.* [15] model the depth of background pixels with a mixture of Gaussians whose number is pixel dependent. Clapes *et al.* [5] had a different approach; they combine RGB and depth information in a four-dimensional vector and then model the background with four-dimensional Gaussians. The initialization is done by means of a time-windowed Gaussian for a static scene, however the model is not adapted during later steps. Camplani and Salgado [4] propose a slight modification for the case of the Kinect camera. Their algorithm uses a mixture of Gaussians to segment the depth map but the number of Gaussians that model the background is controlled by a parameter such that it removes a bias due to the depth uncertainty of the Kinect. This adaptation was also used in [3].

To the best of our knowledge, there exists only one algorithm for the segmentation of the foreground, developed specifically for depth images. This algorithm, proposed by del-Blanco *et al.* [7], is designed for the Kinect camera and it combines two classifiers to produce the foreground/background segmentation map. The first classifier, which is the mixture of Gaussians proposed by Camplani *et al.* [4] applied to depth, serves to provide a pixel-based model for dynamic backgrounds. The second classifier is a Bayesian network that operates at the level of regions to deal with spatial and depth correlations of regions over time. It runs as follows. A dynamic spatial model estimates which regions could belong to the foreground in the current frame, given a previous time. A second, based on depth, predicts the distribution of depth changes between two consecutive frames. While this combination is interesting, these two models rely on restrictive hypotheses. The dynamic spatial model ignores the displacement vector of moving objects, their speed as well as their acceleration, and assumes an isotropic displacement distribution. The model based on depth assumes a unique displacement for all foreground objects, thus ignoring possibly opposite displacements, object deformations, and neglecting that pixels of an object could have different motion speeds. Despite all these limitations, their results show an improvement with respect to some state-of-the-art techniques developed for color images [1, 11, 16, 23]. Unfortunately, the algorithm of del-Blanco *et al.* is far from being real-time; the authors mention a frame rate of about 1 fps on a 3GHz Intel Core i7-3540M for 640x320 large images. There is thus the need for fast algorithms suited for depth images.

3. DESCRIPTION OF OUR BACKGROUND SUBTRACTION METHOD

In this section, we propose a new pixel-based model for background subtraction in 3D images. This model, developed for dealing with depth information acquired with the Kinect version 1, is a hybrid model that combines depth values gathered in one model and wrong values, considered as holes, in another model. First, we explain our motivation for a hybrid model (Section 3.1). Then, we detail both parts of this model (Sections 3.2 and 3.3). Finally, a special post-processing filter that preserves a high reliability over time is presented in Section 3.4.

3.1. Towards a hybrid background model

The choice of a hybrid background model is motivated by the existence of depth measurement failures, for certain pixels in 3D images, often referred to as "holes" in the literature. These failures originate from several possible physical phenomena: multiple reflections, strong light in the scene, depth shadows, absorption by black objects, diffraction, depth discontinuities, etc. When a depth measure is unreliable, the Kinect camera identifies it by a special value in its output; other cameras might have another convention. Obviously, color-based motion detection methods are unaware of the existence of such a special value as the color modality is not affected by holes. Therefore, when a color-based method is applied to range images, the special value is treated as a depth value. This careless treatment of wrong pixel values impacts on the quality of the background model, and the background then mixes real depth values and holes, as illustrated in Fig. 1. In particular, we show the color image of the scene, the depth map given by the Kinect camera, and the background provided by the Pfinder method [22] in, respectively, Figs 1a, 1b, and 1c.

To avoid the appearance of wrong values originating from holes in the background, our technique deals with depth data and holes separately, and it builds two independent but complementary models. One model represents the depth of the scene with a single depth value; the other model analyzes the dynamics of holes and locate pixels whose background should be labeled as a hole. It appears that, for building a reliable background model and subsequently to improve the detection of motion, it is interesting to explicitly identify pixels whose depth measures are never performed (for instance because the real depth of these pixels exceeds the depth range of the sensor) and to consider them as holes rather than to allocate them a depth measure.

Therefore our technique is based on a hybrid background model of the scene that identifies pixels as a depth measure or as a hole, depending on their location. Fig. 1d illustrates the background obtained with our model. The comparison with the background model of the Pfinder model [22] shown on Fig. 1c confirms the relevance of a double model, one for

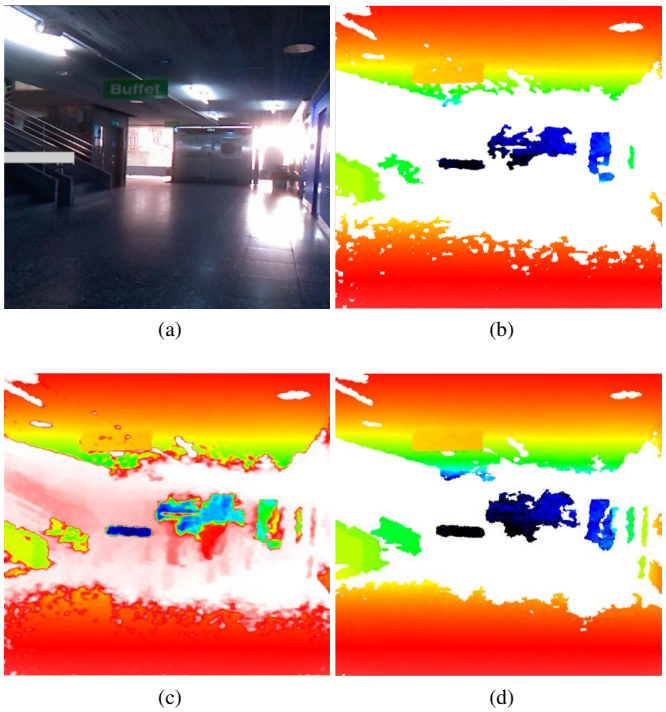


Fig. 1: Towards a hybrid model for representing the background of a range image. The (a) image is the color image of a scene as taken from the database of [19]. Image (b) is the corresponding depth map of the scene (also taken from the database of [19]). In this image, values corresponding to holes are mentioned in white. Image (c) shows the background model obtained with Pfinder [22]. This model wrongly interprets holes in the background as being very close to the Kinect (see the pink pixels). Image (d) presents the background built by our hybrid background model. Note that, in our model, we preserve the distinction between real depth values and holes in the background.

depth and one for holes.

Note that our approach significantly differs from existing techniques and in particular from the one proposed by del-Blanco *et al.* [7]. In their paper, they use a Bayesian network to predict the 3D localization of foreground objects whereas our algorithm models the background. In the next two sections, we describe our model for holes (Section 3.2) and our model for real depth values (Section 3.3).

3.2. Considering holes in one model

The goal of introducing a model for holes in a background subtraction technique applicable to range images is the localization of pixels whose background is better represented by holes instead of a depth. More precisely, we want to identify locations for which depth measures are never performed when the background is not occluded by a foreground object. We will refer to these pixels as *constant holes* in the rest of the paper. Under the assumption that the background is visible most of the time, the pixels that we try to locate correspond to large series of consecutive holes in the video stream. A simple way to build the model then consists in using a counter for each pixel, memorizing the number of consecutive holes. If the counter exceeds a given threshold N_H, the pixel is labeled as a constant hole in the final hybrid background model. For such a pixel, a valid depth measure can only be due to a moving object occluding the background. The classification process is thus the following: if a depth measure is valid, the pixel is classified as foreground, otherwise it is classified as background.

This counting procedure cannot work if the background is not visible most of the time. For instance, in the case of a scene with many moving objects, constant holes may not be identified as such, leading to them being put in the depth-based background model in the final hybrid model. Note that this does not mean that the segmentation algorithm will fail to detect moving objects reliably in this particular situation; it all depends if the depth-based background model is capable to estimate the real depth of the background with a good accuracy.

In order to give an adaptive behavior to our model for holes, we introduce an additional parameter T_W (both T_W and N_H are defined heuristically with $T_W \gg N_H$) used to reset a pixel labeled as a constant hole when necessary. If the number of consecutive holes in a pixel labeled as a constant hole is smaller than N_H during T_W consecutive frames, the pixel is not considered as a constant hole anymore and the pixel model is reinitialized which means that the final model for that pixel is handled by the depth-based model of the hybrid model. To our knowledge, our algorithm is the first algorithm to introduce an analysis of the dynamics of holes in the building of the background model.

3.3. Depth-based background model

The goal of the depth-based background model is to estimate the real depth of the background for each pixel. It is a pixel-based background model which means that for each pixel, a background model is built independently of other pixels. As discussed in [21], this approach should be preferred to region-based models because the segmentation process is local. We thus relegate to post-processing steps the task of adding some form of spatial consistency to our results.

For describing the statistical distribution of depth, we choose a Gaussian probability density function (pdf). Given the predefined form of the pdf, our depth-based model thus belongs to the class of *parametric* models. Compared to *non-parametric* models such as those presented in [11, 21], this class has the advantage of a reduced memory footprint as it only requires to memorize the parameters of the pdf instead of a collection of previously observed values.

For the color modality, a simple model such as the Gaussian is generally not sufficient because the real pdf may have several modes (due to monitor flicker, swaying trees, waves in the sea, ...). In range images however, color changes in the background don't affect the depth measure. Moreover, since the Kinect camera is an indoor device, it can be reasonably assumed that the scene doesn't contain aperiodic dynamic backgrounds such as swaying trees or waves or that these elements are too far to be detected. The choice of one unimodal Gaussian model per pixel is therefore appropriate to represent the measured depth. Consequently, two parameters are memorized for each pixel: the mean of the Gaussian μ and its standard deviation σ. In its simplified form, the BG/FG classification process for the depth-based model considers that the current depth value D_t belongs to the Gaussian distribution if the condition $\mu - K\sigma \leq D_t \leq \mu + K\sigma$ is met, where K is a global parameter controlling the rate of pixels incorrectly classified as foreground in the segmentation map.

From our experience, it is important to improve this decision process. In the next sections, we first elaborate on the nature of the depth signal and its impact on the process. Then, we present the updating equations of the mean μ and the standard deviation σ. We will see that μ is updated by means of a physical interpretation of the depth signal which is the major contribution of this paper, whereas the standard deviation σ is updated according to a law defined by the noise of the sensor. We also propose a mechanism for a fast suppression of ghosts that exploits the physical meaning of the depth signal and a kinematic constraint on the speed of foreground objects (see subsection 3.3.3). The whole segmentation process for the depth-based model is summarized in subsection 3.3.4.

3.3.1. Physical interpretation of the depth signal

The main innovation of our algorithm lies in the physically motivated updating strategy of the Gaussian mean μ. The idea is that depth values can be ordered to the contrary of colors. To illustrate this idea, let's imagine the following problem. Consider a noise-free color image and suppose, to keep it simple, that for a given pixel, we have observed only two different colors: red and blue. We know that one of them is the background color whereas the other is the color of a moving object observed previously for that pixel. The goal consists to identify the color of the background. Given the red or blue information only, it is impossible to answer the question because there is no ordering relationship for colors. The case of range images is different. For a certain pixel, we have two different noise-free depth measures D_1 and D_2 with $D_1 < D_2$. One of these values corresponds to the background whereas the other is the depth of a foreground object observed previously for that pixel location. Here too the goal consists to determine the background value. But now, we can exploit the ordering relationship of range images: the background is always located behind the foreground, by definition. Consequently, D_2 is the depth of the background. This ordering relationship is valid for an arbitrary number of values. If we have n noise-free values $D_1, D_2, \ldots, D_n$ such that $D_1 < D_2 < ... < D_n$ with one value only belonging to the background, then D_n is its depth value. The above physical interpretation allows us to conclude that for a noise-free signal –please note that is a strong condition–, the depth of the background at a specific pixel location is at least equal to the maximum depth value observed for that pixel. We can thus approximate the mean of the Gaussian by $\mu_t = MAX(D_k)$ for $k \in [0,\, t]$ where D_k is the measured depth at time k.

This physical interpretation has a huge benefit. It allows us to suppress ghosts that would appear on color images when a static object belonging to the background suddenly starts moving. If a background initially masked by this object is uncovered, it is erroneously classified as foreground by color-based methods because of the absence of an ordering relationship to discriminate this case from the case of a foreground object masking the background. The Sakbot system of Cucchiara *et al.* [6] eliminates these ghosts by means of an optical flow computation on the foreground connected components of the segmentation map. This highly time-consuming process is unable to suppress ghosts connected with the moved object. ViBe [21] provides a suppression of ghosts faster than the incorporation of stopped objects into the background thanks to a conservative updating strategy and a spatial diffusion mechanism of background pixel values. However, the randomness of the spatial diffusion and its limited range between two consecutive frames (3×3 square grid) prevent ghosts from being eliminated quickly. By contrast, with our technique, the uncovered background will be instantaneously incorporated into the background model as it is the maximum observed value and the ghost will thus be eliminated at the next frame. Fig. 2 illustrates this enhancement.

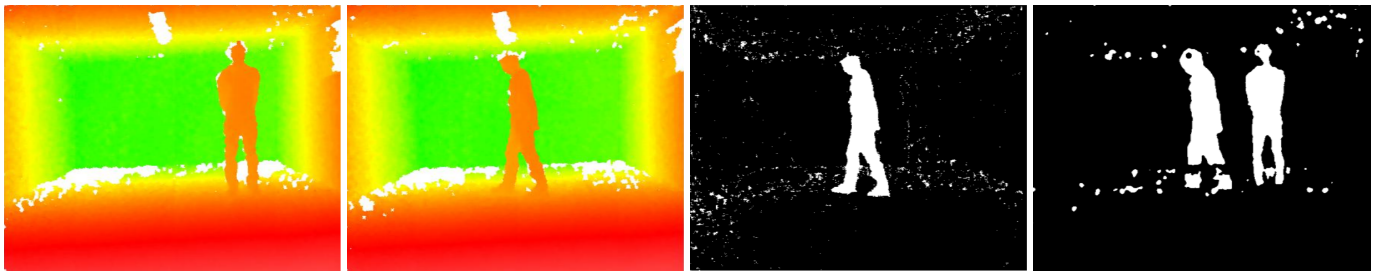
3.3.2. Depth-dependent BG/FG decision threshold

We have just explained how the physical nature of depth signal affects the updating mechanism for the mean. The updating procedure for the standard deviation is instead based on a behavior law of the sensor noise. Khoshelham [14] established that the noise of the Kinect depth sensor follows a quadratic relationship with respect to the actual depth of the object:

$$\sigma_d = K_{kinect} D^2, \tag{1}$$

where σ_d is the standard deviation of depth, D the real depth of the surface and K_{kinect} a constant depending on the reflecting properties of the material. For a planar surface nearly perpendicular to the optical axis of the camera, $K_{kinect} \approx 1,5.10^{-3} m^{-1}$ (this is the value taken in this paper). We exploit this result in our depth-based background model by stating that:

$$\sigma_t = K_{kinect} \mu_t^2. \tag{2}$$



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

Fig. 2: How the ordering relationship between range values helps to remove ghosts almost instantaneously. (a) and (b) are depth maps of a scene showing a person that starts moving. In (b), the person has moved, leaving an uncovered background behind her. Image (c) is the segmentation map produced by our method at the time of the (b) image; there is no ghost to be seen to the contrary of classical techniques, as shown by the segmentation map (d), obtained with the PBAS technique [11].

The standard deviation σ_t is thus a good estimate of the measurement uncertainty if the mean μ_t is close to the actual depth of the background. As stated earlier, a current pixel depth value D_t fits within the background model if the following condition

$$\mu_t - K\sigma_t \leq D_t \leq \mu_t + K\sigma_t \tag{3}$$

is met. Therefore, our BG/FG decision threshold is given by $\tau_t = KK_{kinect}\mu_t^2$. This equation shows that our BG/FG decision threshold is depth-dependent, making our method able to adapt itself to the non-uniform noise in range images. Unlike traditional methods that would use a constant decision threshold for the whole image or the full depth range, our technique can deal both with low or large depth values, as shown in Fig. 3.

3.3.3. Kinematic constraint and faster suppression of ghosts

As illustrated in Fig. 2, the updating of the μ parameter of a pixel by means of the maximum observed depth value is the key for the fast suppression of ghosts. However, this suppression is only effective after one frame. Indeed, when the scene background is uncovered, the depth-based background model is updated. However, we still have to present our classification criterion for this case.

This process involves a constraint on the speed of the moving objects. The knowledge of a physical limit on the speed of the foreground objects can be beneficial for the task of motion detection. If we denote this maximum speed by V_{max}, the maximum depth jump of the foreground between two consecutive frames is upper bounded by:

$$\triangle P_{max} = \frac{V_{max}}{Fr}, \tag{4}$$

where Fr is the frame rate of the camera (30 frames per second for the Kinect). Therefore, if the current depth value D_t is such that

$$\mu_t + K\sigma_t + \triangle P_{max} < D_t, \tag{5}$$

then we can reasonably assume that the depth jump cannot be due to the backwards motion of a foreground object. The background label is then assigned to the pixel. This implies an immediate suppression of ghosts. However, if the following condition is met:

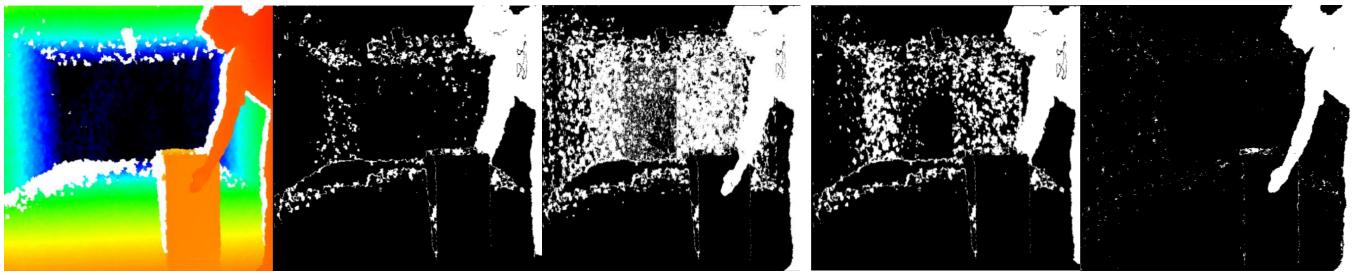
$$\mu_t + K\sigma_t < D_t \leq \mu_t + K\sigma_t + \triangle P_{max}, \tag{6}$$

the depth jump can be associated either to the backwards motion of a foreground object or to an uncovered background. In this case, as depth jumps due to an uncovered background generally follow the inequality (5), we interpret the situation as a backwards motion of a foreground object and assign a foreground label to the pixel.

3.3.4. Summary of the updating equations, classification process, and initialization process

Table 1 summarizes the updating equations, the classification process and the initialization process of our depth-based background model. Note that when the current depth value fits within the background model, we use a recursive filter on the mean μ to enhance the estimation of the real background depth. The parameter α is a global parameter set for the whole video and satisfies $0 \leq \alpha \leq 1$. The fitting interval between the background model and the current depth value is thus the only interval for which the mean μ can decrease.

When a foreground object moves in front of the background, parameters are not updated. This means that foreground objects that suddenly stop are not incorporated into the background. As said in [21], this is preferable as the application may need to keep those objects in the foreground. We thus relegate to post-processing steps the decision to keep newly static objects in the foreground or to absorb them in the background.



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

Fig. 3: A depth-dependent BG/FG decision threshold produces high quality BG/FG segmentation for both low and large depth values. Image (a) shows a depth map with depth camouflage for low depth values (hand of the man). We compare our result with those of the Sakbot system presented in [6]. This system uses a global BG/FG decision threshold for the whole image. Images (b), (c) and (d) show Sakbot segmentation results for a high threshold, a low threshold, and a medium threshold respectively. None of them produces a reliable segmentation for both low and large depth values. By contrast, (e) shows that our method provides a high quality segmentation.

Updating equations and classification process					
	0 — L_t — μ_t ($K\sigma_t$) — H_t — $H_t + \Delta P_{max}$ → *Depth*				
Condition	$D_t = 0$ (hole)	$0 < D_t < L_t$	$L_t \leq D_t \leq H_t$	$H_t < D_t \leq H_t + \Delta P_{max}$	$H_t + \Delta P_{max} < D_t$
μ_{t+1}	μ_t	μ_t	$(1-\alpha)\mu_t + \alpha D_t$	D_t	D_t
σ_{t+1}	σ_t	σ_t	$K_{kinect}\mu_{t+1}^2$	$K_{kinect}\mu_{t+1}^2$	$K_{kinect}\mu_{t+1}^2$
Class	BG	FG	BG	FG	BG
Initialization process					
$\mu_0 = D_0$				$\sigma_0 = K_{kinect}\mu_0^2$	

Table 1: Summary of the updating equations, classification process, and initialization process of the depth-based background model. D_t is the current depth value for the considered pixel. L_t and H_t are respectively defined by $\mu_t - K\sigma_t$ and $\mu_t + K\sigma_t$.

The updating strategy presented on table 1 may be seen as a *semi-conservative* strategy because pixel values that do not fit within the background model may be used to update the model or not, depending on the position of these values with respect to the $[L_t,\ H_t]$ interval.

3.4. Post-processing

In order to provide a high quality foreground segmentation over time, a post-processing filter is applied to cope with an event that is ignored by our depth-based background. In fact, the model assumes that all noisy measurements D_t of the background depth satisfy $\mu_t - K\sigma_t \leq D_t \leq \mu_t + K\sigma_t$. However, when due to noise, a depth exceeds the upper limit of this inequality, the mean μ is set equal to this outlier value. Next depth values, which are not outliers, will met the $0 < D_t < L_t$ condition of table 1, which means μ is not updated anymore, leading to a persistent error. This increases the amount of false positives in the segmentation map, referred to as "outliers errors" in the rest of the section.

To cope with this effect, we introduce a mechanism whose task consists to lower the mean value μ of the model after some outliers. This is not trivial because the connected components of both outliers errors and a sleeping foreground are static in the segmentation map. Therefore, counting the number of consecutive foreground labels or using an optical flow computation on the foreground objects is ineffective to distinguish outliers errors from a sleeping foreground.

To make the distinction, we rely on the *contour* of connected foreground pixels in the segmentation map. As range measurements near depth discontinuities oscillate between the depth values of both sides of the discontinuity, the contour of a sleeping foreground object oscillate too. On the contrary, the contour of a connected component of outliers errors is stable because it is not linked to any depth discontinuity. The proposed post-process consists in counting, for each pixel, the number of consecutive motion maps for which the pixel is classified as foreground and located on the contour of a connected foreground component. If this number exceeds a global parameter N_{max}, the mean μ is reset by the equation $\mu_{t+1} = D_t$ with a probability $p = \frac{1}{L}$, where L is the length (number of pixels) of the contour containing the considered pixel. The introduction of this probability results from a compromise between the correction of outliers errors and the inhibition of the incorporation of foreground objects into the background. It is a heuristic.

Finally, as commonly adopted in background subtraction,

we add two post-processing operations: a morphological opening with a 3×3 cross structuring element followed by a 7×7 median filter. The morphological opening mainly deletes false positive errors caused by the multi-modal behavior of the depth pdf near jump edges. The median filter adds some form of spatial consistency to the results of our pixel-based technique.

4. EXPERIMENTAL RESULTS

In this section, we compare the results of our technique with those of two state-of-the-art color-based methods (PBAS [11] and SOBS [16]) and those of two well-known Gaussian mixtures: the mixture of Stauffer *et al.* [20] and that of Zivkovic *et al.* [23]. We used the implementations available in the BGSLibrary [18] and set the parameters of these methods to the values recommended by the authors in their papers. Regarding the parameters of our method, the following set has been used: $\alpha = 0.01$, $K = 3.5$, $V_{max} = 30\,km/h$, $N_{max} = 10$, $N_H = 80$, $T_W = 300$.

As there is no available diversified database usable for the evaluation of motion detection in range images, we have built a new database that contains eight depth maps sequences. Three sequences have been taken from an existing depth-based database [19], representing a real surveillance application. We have added five sequences representing various challenges: a basic scene with a person moving continuously, a sequence with depth camouflage, another with many foreground objects in the initialization phase, a sequence evaluating the ability of the algorithm to distinguish black moving objects from depth shadows, and a video containing a person initially at rest that suddenly starts moving laterally and then stops. This last video is designed to evaluate the ability of the algorithm to suppress ghosts and to keep sleeping objects in the foreground. For all these sequences, ground-truths have been labeled manually at the rate of one ground-truth image per 25 frames.

The metric used to evaluate performances is the Euclidean distance in the ROC space between the position of the binary classifier and the position of the best theoretical classifier. This metric is thus defined by $d_{euc} = \sqrt{FNR^2 + FPR^2}$, where FNR and FPR are respectively the false negative rate and the false positive rate. Compared to classical metrics such as error rates, recall, specificity or F-measure, the Euclidean distance has the advantages to equally consider false positive and false negative detections, and it is independent of the prior probability for the pixel to belong to the background. The comparative results are drawn in Fig. 4, on the ROC space for each sequence. This figure shows that our proposed method outperforms state-of-the-art methods for 3D images.

Ideally, we would have liked to compare our approach to that proposed by del-Blanco *et al.* [7]. Unfortunately, there is no source code available for this method and a personal implementation is unrealistic given the complexity of the technique.

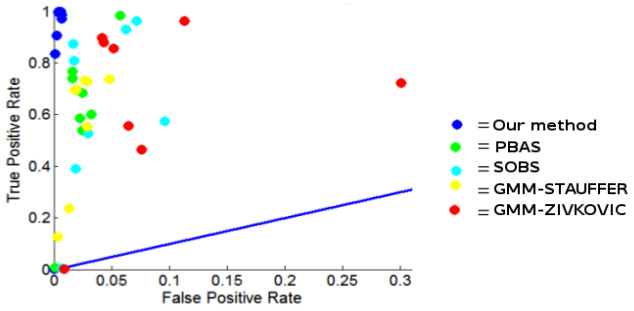


Fig. 4: Comparison of methods in the ROC space for all sequences. Blue line represents the ROC curve of a classifier that guesses the class of the pixel randomly. The results of our method (blue spots) are very close to the position of the best theoretical classifier defined by $(FPR, TPR) = (0, 1)$, to the contrary of other known techniques developed for color images.

However, regarding the computational requirements, our approach is much more effective as it is capable to process 15 frames per second on a 2.4 GHz Intel Pentium 2020M processor for 640x480 large image sizes (in a C++ implementation, with no particular optimization), whereas the Bayesian network runs at about 1 frame per second on a 3 GHz Intel Core i7-3540M for 640x320 large image sizes. Our algorithm is thus suited for real-time applications.

5. CONCLUSION

In this paper, we present a novel pixel-based background subtraction technique for 3D images. The technique is based on a hybrid model that combines two models, one for labeling holes, and one for the estimation of real depth values. The major contribution is the exploitation of the physical meaning of the depth signal in the depth-based background model to improve the detection of foreground objects and to instantaneously suppress ghosts that would appear in color images. The non-uniformity of the spatial distribution of noise in range images is also considered to improve the quality of the segmentation both for low and large depth values. Experiments show that our method outperforms state-of-the-art color-based methods applied to range images.

Acknowledgements.

A. Lejeune is under a contract funded by the FEDER program of the Walloon Region, Belgium.

6. REFERENCES

[1] O. Barnich and M. Van Droogenbroeck. ViBe: A universal background subtraction algorithm for video sequences. *IEEE Trans. Image Process.*, 20(6):1709–1724, June 2011.

[2] T. Bouwmans. Recent advanced statistical background modeling for foreground detection - a systematic survey. *Recent Patents on Computer Science*, 4(3):147–176, Sept. 2011.

[3] M. Camplani, C. del Blanco, L. Salgado, F. Jaureguizar, and N. Garcia. Advanced background modeling with RGB-D sensors through classifiers combination and inter-frame foreground prediction. *Machine Vision and Applications*, 25(5):1197–1210, July 2014.

[4] M. Camplani and L. Salgado. Background foreground segmentation with RGB-D Kinect data: an efficient combination of classifiers. *J. of Visual Communication and Image Representation*, 25(1):122–136, Jan. 2014.

[5] A. Clapes, M. Reyes, and S. Escalera. Multi-modal user identification and object recognition surveillance system. *Pattern Recognition Letters*, 34(7):799–808, May 2013.

[6] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(10):1337–1342, Oct. 2003.

[7] C. del Blanco, T. Mantecón, M. Camplani, F. Jaureguizar, L. Salgado, and N. García. Foreground segmentation in depth imagery using depth and spatial dynamic models for video surveillance applications. *Sensors*, 14(2):1961–1987, Feb. 2014.

[8] B. Freedman, A. Shpunt, M. Machline, and Y. Arieli. Depth mapping using projected patterns, 2010. US Patent Application 20100118123.

[9] K. Greff, A. Brandão, S. Krauß, D. Stricker, and E. Clua. A comparison between background subtraction algorithms using a consumer depth camera. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 1, pages 431–436, Rome, Italy, 2012.

[10] S. Guðmundsson, R. Larsen, H. Aanæs, M. Pardás, and J. Casas. TOF imaging in smart room environments towards improved people tracking. In *IEEE Comp. Society Conf. on Comp. Vision and Pattern Recognition Workshops*, pages 1–6, Anchorage, AK, USA, June 2008.

[11] M. Hofmann, P. Tiefenbacher, and G. Rigoll. Background segmentation with feedback: The pixel-based adaptive segmenter. In *IEEE Int. Conf. Computer Vision and Pattern Recognition Workshop (CVPRW)*, Providence, Rhode Island, June 2012.

[12] B. Jansen, F. Temmermans, and R. Deklerck. 3D human pose recognition for home monitoring of elderly. *International Conference of the Engineering in Medicine and Biology Society*, pages 4049–4051, Aug. 2007.

[13] P.-M. Jodoin, S. Piérard, Y. Wang, and M. Van Droogenbroeck. Overview and benchmarking of motion detection methods. In T. Bouwmans, F. Porikli, B. Höferlin, and A. Vacavant, editors, *Background Modeling and Foreground Detection for Video Surveillance*, chapter 24. Chapman and Hall/CRC, July 2014.

[14] K. Khoshelham. Accuracy analysis of Kinect depth data. In *ISPRS Workshop Laser Scanning 2011*, volume XXXVIII-5/W12, Calgary, Canada, Aug. 2011.

[15] J. Lorenzo-Navarro, M. Castrillon-Santana, and D. Hernandez-Sosa. On the use of simple geometric descriptors provided by RGB-D sensors for re-identification. *Sensors*, 13(7):8222–8238, June 2013.

[16] L. Maddalena and A. Petrosino. A self-organizing approach to background subtraction for visual surveillance applications. *IEEE Trans. Image Process.*, 17(7):1168–1177, July 2008.

[17] L. Schwarz, A. Mkhitaryan, D. Mateus, and N. Navab. Human skeleton tracking from depth data using geodesic distances and optical flow. *Image and Vision Computing*, 30(3):217–226, Mar. 2012.

[18] A. Sobral. BGSLibrary: An OpenCV C++ background subtraction library. In *Workshop de Visao Computacional (WVC)*, Rio de Janeiro, Brazil, June 2013.

[19] L. Spinello and K. Arras. People detection in RGB-D data. In *IEEE/RSJ Int. Conf. Intelligent Robots and Syst. (IROS)*, pages 3838–3843, San Francisco, USA, Sept. 2011.

[20] C. Stauffer and E. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 246–252, Ft. Collins, USA, June 1999.

[21] M. Van Droogenbroeck and O. Barnich. ViBe: A disruptive method for background subtraction. In T. Bouwmans, F. Porikli, B. Höferlin, and A. Vacavant, editors, *Background Modeling and Foreground Detection for Video Surveillance*, chapter 7, pages 7.1–7.23. Chapman and Hall/CRC, July 2014.

[22] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):780–785, July 1997.

[23] Z. Zivkovic and F. van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, 27(7):773–780, May 2006.