# Techniques to improve the foreground segmentation with a 3D camera and a color camera

Sébastien PIÉRARD, Jérôme LEENS, Marc VAN DROOGENBROECK
{Sebastien.Pierard, Jerome.Leens, M.VanDroogenbroeck}@ulg.ac.be
INTELSIG, Laboratory for Signal and Image Exploitation
Montefiore Institute, University of Liège, Belgium

***Abstract*—Nowadays, techniques for real-time interpretation of video scenes are widespread. Amongst these techniques, the foreground segmentation is one of the favorite. It can be applied to color images as well as depth maps. The point of using depth maps is straightforward as a single color camera is not able to provide depth information. Technologies capable to acquire 3D informations are thus adequate to complement color cameras in consumer products. Practice has shown that 3D or RGB signals, taken alone, are unreliable to extract the foreground under arbitrary conditions. Therefore we combine both modalities to counter the intrinsic limitations of both modalities, which is only possible if the problems specific to a technology are handled appropriately. This paper presents a new global approach for enhanced foreground segmentation that handles limitations to 3D and RGB in a combined way.**

***Index Terms*—Video interpretation, 3D camera, depth camera, range camera, background subtraction.**

## I. INTRODUCTION

One of the most challenging tasks in computer vision is the real-time interpretation of scenes. This topic has many applications including video-surveillance for security or safety, man-machine interaction, immersive games, etc.

In applications using cameras for acquisition, a widespread technique used as a pre-processing step for interpretation is the foreground segmentation (also called background subtraction). This technique separates pixels of the background, where no motion is detected, from pixels of moving objects contained in the foreground. Foreground objects correspond either to the users, or to the physical objects they interact with, located in the foreground of the scene. Background subtraction is important in that it provides the user a zone of interest related to motion and shapes, instead of static textures, and thus decreases the sensitivity to appearance. In addition, robust techniques, such as ViBe [1], have proved to be resilient to important amounts of noise. Having a motionless camera and stationary (or controlled) lighting conditions are the only constraints imposed by most of background subtraction algorithms.

The use of background subtraction techniques is not limited to signals acquired by grayscale or color cameras. Background subtraction can also be used on depth maps, but it requires some caution because 3D maps are subject to more noise and their resolution is significantly lower than that of conventional cameras; we need to select a background subtraction technique that performs well even with low resolution images.

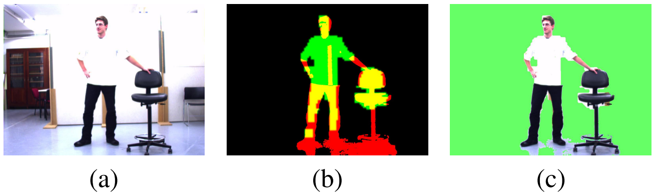

(a) (b) (c)

Figure 1. The goal of foreground segmentation is to isolate the users and the objects they interact with. (a) A color image taken from a video. (b) The superimposition of the foreground segmentation computed from the color modality (in the red channel) and the one computed from the 3D modality (in the green channel). Note that none of the 3D or RGB signals are reliable to extract the foreground, but that combining both modalities improves the segmentation because the two signals have a different physical significance. (c) The segmented foreground.

Technologies that acquire 3D maps are useful because they offer a depth information that color cameras do not measure and, without depth information, even simple actions are difficult to recognize. As of today, there are three families of technologies to get 3D informations:

1) Stereo-vision require a nontrivial processing to precisely register points of two images. The lack of texture can crumble this system.
2) Asking users to wear some sensors on their body parts to be tracked is another way to capture 3D informations. Some examples of this, amongst others, are systems based on *Motion Capture Unit*s, and controller-based systems such as the *Wiimote*. However, these technologies does not provide depth maps.
3) Finally, there are depth-cameras (also called 3D cameras), directly measuring depth maps, and which are not intrusive and use their own source of light. With these technologies, body motion can be captured without any controller. *Microsoft* has recently announced a new interface (*Project Natal*) based on this technology for the next generation of its *Xbox* product.

In this paper, we concentrate on using an RGB and a 3D camera simultaneously to segment the foreground. The point in using an RGB camera is to enhance the low resolution of 3D cameras. Moreover, as explained in [2], combining the foreground segmentations of color images and depth maps usually improves the segmentation when the use of a sole modality fails to build a satisfactory segmentation (see Fig. 1). Unfortunately, combining the two modalities has not only advantages, but also disadvantages. This paper concentrate on

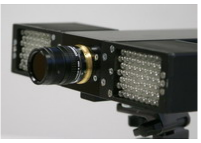


Figure 2. A 3D camera (by PMD Technologies). Two arrays of infrared LEDs are located on both sides of the sensor.

explaining these disadvantages and presenting algorithms to solve them.

This paper is organized as follows. Our experimental setup and the image registration process are described in Section II. The main principles of the depth camera technology that we use, and the related drawbacks are presented in Section III. Section IV describes the foreground segmentation for the color and depth modalities, and how to combine the segmentation maps. Section V explains the drawbacks proper to each modality and how the other channel permits to solve the respective issues. Section VI concludes the paper.

## II. Image acquisition and registration

### A. Our experimental setup

We use an RGB camera and a 3D camera attached on both side of a horizontal plane surface. Both cameras are equipped with identical lenses, but their fields of view are different. They are as close as possible to each other, to reduce occlusion problems. All the discussions of this paper relate to the use of indoor PMD (Photonic Mixer Device) cameras. The device we used is a *PMD[vision]19k* (Fig. 2) , which acquires low-resolution depth maps (160 × 120 pixels). Such a camera illuminates the scene with its own infrared light source, and can thus operate in a total darkness. Nevertheless, we use visible lighting to fulfill the requirements of the RGB modality.

### B. Image registration

Because the two modalities are acquired with different sensors, a registration of the two images is required to combine the foreground segmentations.

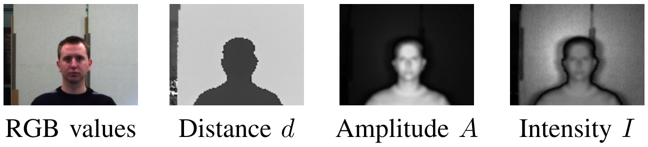
The major difficulty for registration is that it depends on the content of the scene. Ideally, one would have a dynamic mapping between the PMD and the RGB cameras based on the depth map. Although the calibration of cameras is a classical problem in computer vision, the calibration of the PMD camera is much more complicated than the calibration of the RGB camera, because of its low resolution. Therefore, we want to skip a calibration process, and we have chosen to establish a direct transform, which is based on three assumptions:

1) the two cameras follow the pinhole model;
2) the optical axes are parallel;
3) the depth difference between the two optical centers is negligible compared with the scene depth.

These assumptions lead to the following mapping between the PMD pixel $(u, v)$ and the RGB pixel $(u', v')$ :

$$z(u,v) = d(u,v)\ \left(1 + (Au+B)^2 + (Cv+D)^2\right)^{-1/2}$$

$$u'(u,v) = Eu + Fv + G + H\ z^{-1}(u,v)$$



RGB values Distance $d$ Amplitude $A$ Intensity $I$

Figure 3. A color image and the 3 channels provided by a 3D camera.

$$v'(u,v) = Iu + Jv + K + L\ z^{-1}(u,v)$$

The capital letters stand for constants, $z$ is the depth and $d$ is the measured distance (the proof of this new result is not given here). The main drawback of this dynamic mapping is the influence of the $d$ channel, which is noisy. Our experiments showed that a spatial gaussian filtering of the channel $d$ with $\sigma \geq 2$ pixels is necessary.

Note that the objective of this transform is to align the foregrounds of the 3D and RGB channels. The advantage of our transform is that the background has no impact on the segmentation of the foreground so that we can concentrate on the foreground only. Assuming that only one user is present in the scene and that this user stands at a distance of at least 2 meters, it is reasonable to assume further that his distance to the camera is unique (this assumption is made for the mapping only!). In this case, it is possible to use a static mapping (an affine transformation) instead of a dynamic mapping. To minimize the registration distance error, in the least-squares sense, the depth $z$ used for determining the affine transformation coefficients is the harmonic mean of the depths over the foreground.

## III. PMD cameras and their drawbacks

PMD cameras illuminate the scene with their own amplitude modulated infrared light source. The sensors provide three channels per pixel :

1) the *distance* $d$ is proportional to the phase shift between the emitted signal and the received one. $d$ is an estimation of the distance between the camera and the corresponding point of the scene;
2) the *amplitude* $A$ is proportional to the amplitude of the alternating component (AC) of the received signal. It measures the strength of the signal used to compute $d$. It is thus an indicator about the accuracy of the distance estimation;
3) and the *intensity* $I$ is proportional to the direct component (DC) of the received signal. It reveals the luminance of the scene.

The three channels given by a PMD camera are imprecise (see Fig. 3). The channel $d$ is not reliable for many reasons, some of them being that:

- $d$ is corrupted by a lot of noise and is neither an accurate nor a precise distance estimator;
- $d$ depends on the orientation and surface of the reflecting materials;
- the distance estimation is biased if the object is too close to the camera (saturation of the sensor) or if it is too far from the camera (ambiguity after 7.5 m for a 20 MHz modulation);

- because of persistence effects, when a fast movement occurs, a trail can be observed on the three channels of the PMD camera. If the observed person moves quickly, it will appear twice in the images, in its starting position, and in its ending position.

The $A$ and $I$ channels have similar limitations (saturation for close objects, measures related to physical properties of the observed objects, persistence effect between successive frames), and a shadow effect originating from the lateral position of the infrared sources (see Fig. 2).

Some authors [3] have tried to extract more reliable informations out of the three channels provided by the camera. The three channels are correlated : they are all affected by the properties of the observed surfaces, and all of them depend on the nominal distance (the power attenuation of $A$ and $I$ depends on the nominal distance). However, correcting the values of the channels provided by the PMD camera is impossible in uncontrolled scenes, because both the channels $A$ and $I$ depend on unknown properties of the scenes.

## IV. Foreground segmentation

We now describe how to build a segmentation map for each modality. We also enumerate the various kinds of errors affecting the segmentations. Those ones are classified in accordance with their effects on the pixel classification:

1) pixels that are erroneously classified in the foreground;
2) and pixels that are erroneously put in the background.

Combining the two modalities enable us to compensate for the second class. On the other hand, our combination algorithm has drawbacks: the errors of the first class are cumulative. This section explains how to get around these errors.

### A. Background subtraction on color images

Color cameras usually have a high resolution, compared to 3D cameras. This results in high precision silhouettes. ViBe is applied on the luminance ($Y$ channel) to build the *y-foreground*. A morphological opening is used to remove noise. An alternative would be to apply ViBe on the three RGB channels, and to combine the segmentation maps to build an *rgb-foreground*. However, the improvements do not justify the computational overhead.

With this modality, some pixels could be erroneously classified in the foreground: (i) if the lighting conditions are not stationary; (ii) if there are shadows; (iii) if the camera is moving; or (iv) if the background is not static.

Other pixels could also be misclassified in the background: (i) if the colors of the users match those of the background ; or (ii) if the lighting conditions are too low.

### B. Background subtraction on depth maps

ViBe is simultaneously applied to the three channels given by the PMD camera. The three resulting segmentation maps are then combined to build the *pmd-foreground*. A pixel belongs to the *pmd-foreground* if it belongs to one segmentation map at least. The *pmd-foreground* is filtered by a morphological opening to remove noise. Filling holes in the connected components is also possible.

With this modality, some pixels could be erroneously classified in the foreground : (i) if there is a fast movement, because of persistence effects ; (ii) if the observed object is near to the camera and the background is far enough, because of shadows ; (iii) if the camera is moving ; or (iv) if the background is not static.

Some pixels could also be misclassified in the background : (i) if the user stands at the same distance from the camera than the background ; (ii) if the background is located at more than 7.5 m from the camera ; (iii) if the foreground is too close from the camera, because of saturation ; or (iv) if there are materials absorbing infrared.

### C. Combining the two modalities

For indoor applications, it is quite easy to avoid a lot of errors by taking care of the lighting conditions, avoiding dynamic backgrounds, and ensuring that the distance between the cameras and the background does not overstep 7.5 m.

However, even if those sources of error are avoided, the 3D or RGB signals, taken alone, remain unreliable to extract the foreground under arbitrary conditions. The combination of both cameras allows to avoid the intrinsic limitations of both modalities, only if one solves some problems proper to these technologies, like the persistence of motion in PMD signals. Our algorithm incorporates answers to these problems to correct the foreground map.

Our combination algorithm basically consists in assigning the class "foreground" to a pixel if it stands in the *pmd-foreground* or in the *y-foreground*. The global segmentation map has the size of an RGB image. The mapping step described in Section II-B is used to superimpose the *pmd-foreground* on the *y-foreground*. A linear interpolation is used to increase the resolution.

As explained in Leens *et al.* [2], we hope that the problem of misclassified pixels in the background will be annihilated by combining the two modalities. Circumstances in which the user stands at the same distance from the camera than the background, and its colors match those of the background, are exceptional.

The issue of pixels misclassified in the foreground due to shadows in the RGB images has already been discussed in the literature [4], and will not be detailed by this paper. The following section explains how the algorithm can be modified to avoid the pixels erroneously classified in the foreground due to persistence effects and infrared shadows.

## V. Towards an enhanced combination of depth and color modalities

### A. Solution to the infrared shadows issue

Because the infrared emitters are located on the left and on the right of the sensor, the infrared shadows are always on the left and on the right of the foreground objects. Let $z_{FG}$ be the depth of the foreground, $z_{BG}$ the depth of the background, and $\omega$ a constant. The width of the shadows is given by:

$$\omega\left(\frac{1}{z_{FG}}-\frac{1}{z_{BG}}\right)$$

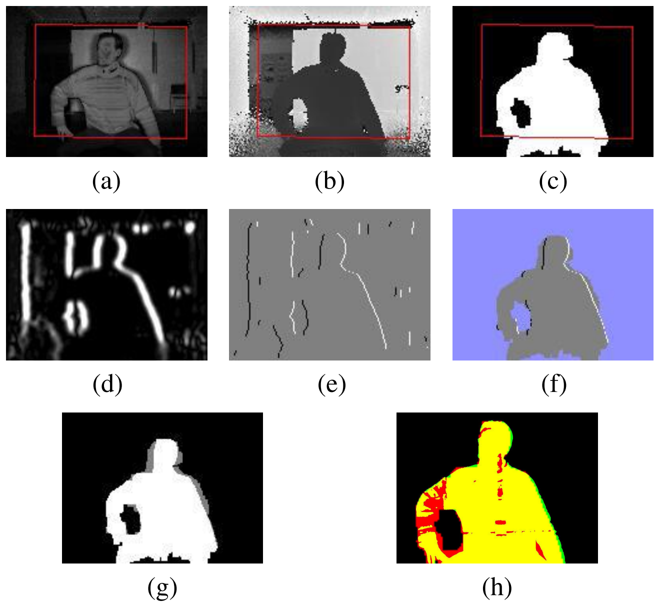


Figure 4. Rectification of the *pmd-foreground* to remove the pixels where there are infrared shadows. The red boxes show the area of the pmd-image that is aligned with the rgb-image.
(a) The intensity channel ($I$) with visible shadows;
(b) The distance channel ($d$). This channel, which is not affected by shadows, is used to correct the *pmd-foreground*;
(c) The *pmd-foreground* prior to its correction;
(d) The absolute value of the horizontal derivative of the channel $d$. A gaussian filter with a standard deviation of 2 pixels is beforehand applied to $d$;
(e) The local maxima (in white) and local minima (in black) of the horizontal derivative of $d$ filtered. A threshold is used to keep only the points corresponding to strong transitions in the channel $d$;
(f) The result of the masking of the image (e) by the silhouette (c). Black points near left edges and white points near right edges indicate where the *pmd-foreground* should stop horizontally after the rectification.
(g) The classification of the pixels. Gray pixels indicate shadows in the *pmd-foreground*. Only the white pixels are kept in *pmd-foreground*;
(h) The comparison of the *rgb-foreground* (in the green channel) and the rectified *pmd-foreground* (in the red channel). The absence of red border indicates that all shadows have been removed, and the absence of green border indicates that we have not removed too many pixels of the *pmd-foreground*.

Thus, it would be possible to predict the width of the shadows if we had a noise-free depth map and if we knew the exact position of the user's contours. However, the dual problem of removing the shadows is much more complicated. First, we don't know where motion detection algorithm cuts in the shadow: the exact position of the contour is thus undetermined. Secondly, the width of shadow is needed to retrieve $z_{FG}$ in the depth map.

Fortunately, the channel $d$ is not affected by shadows, to a first approximation. Thus, the *pmd-foreground* can be rectified by searching strong transitions in $d$. Substantial shadows only exist if the distance between the user and the background is large enough. Thus, in the presence of weak transitions in $d$, there is no need for local corrections. For each pixel of the foreground segmentation contour, we look for strong transitions in $d$ in the neighborhood, on the same line, and inside the foreground. If such a point is found, the contour pixel is moved towards the center of the map. Fig 4 shows the various steps of the algorithm, and the final result.

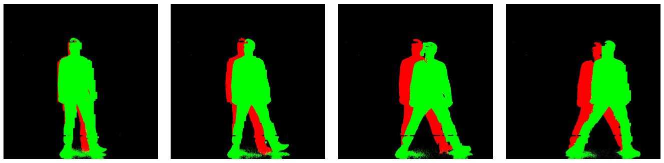


Figure 5. Correction to the persistence effects. Pixels in red are those detected as being affected by persistence. Note that the persistence effect is not negligible.

### B. Solution to the persistence effects issue

Our experiments show that the duration of the persistence effects on the PMD modality is limited to only one frame. Thus, the solution is to compare the *pmd-foreground* and the *y-foreground* computed on the current frame with those computed on the previous frame. A pixel is considered as being affected by the persistence effects if it stays in the *pmd-foreground*, but disappears from the *y-foreground*. If a pixel is affected by the persistence effects, then it is removed from the global segmentation map. Fig. 5 shows the results of this solution.

It should be noted that this algorithm also handles the case where the cameras are not synchronized, if the RGB frame is slightly more recent than the the PMD frame.

## VI. Conclusions

This paper presents techniques to improve the foreground segmentation using both modalities of a 3D camera and a color camera. Combining the signals of both modalities allows to extract a more complete segmentation map in many difficult cases. However, combining the two modalities brings both advantages (pixels correctly recognized as foreground) and drawbacks (pixels erroneously recognized as foreground).

We already discussed about the advantages of combining the RGB and 3D modalities in [2]. This paper focuses on circumventing the disadvantages introduced by the PMD technology: the persistence of motion, and the infrared shadows. Our algorithm offers excellent performances in unconstrained conditions, and runs in real time.

## VII. Acknowledgments

This work has been funded by a *FRIA* grant (Belgium).

## References

[1] O. Barnich and M. Van Droogenbroeck, "ViBe: a powerful random technique to estimate the background in video sequences," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009)*, April 2009, pp. 945–948.

[2] J. Leens, S. Piérard, O. Barnich, M. Van Droogenbroeck, and J.-M. Wagner, "Combining color, depth, and motion for video segmentation," in *Computer Vision Systems, 7th International Conference on Computer Vision Systems, ICVS 2009*, Liège, Belgium, October 2009, pp. 104–113.

[3] M. Lindner and A. Kolb, "Lateral and depth calibration of PMD-distance sensors," in *Advances in Visual Computing*, vol. 2. Springer, 2006, pp. 524–533.

[4] A. Prati, I. Mikic, M. Trivedi, and R. Cucchiara, "Detecting moving shadows: algorithms and evaluation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, July 2003.