Combining Color, Depth, and Motion for Video Segmentation

Jérôme Leens[2] Sébastien Piérard[1*] Olivier Barnich[1]
Marc Van Droogenbroeck[1] Jean-Marc Wagner[2]

ICVS 2009, October 13-15, 2009, Liège, Belgium

[1] INTELSIG Laboratory, Montefiore Institute, University of Liège, Belgium
[2] Haute École de la Province de Liège, Département Ingénieur Industriel, Belgium

Abstract

This paper presents an innovative method to interpret the content of a video scene using a depth camera. Cameras that provide distance instead of color information are part of a promising young technology but they come with many difficulties: noisy signals, small resolution, and ambiguities, to cite a few.

By taking advantage of the robustness to noise of a recent background subtraction algorithm, our method is able to extract useful information from the depth signals. We further enhance the robustness of the algorithm by combining this information with that of an RGB camera. In our experiments, we demonstrate this increased robustness and conclude by showing a practical example of an immersive application taking advantage of our algorithm.

1 Introduction

One of the main tasks in computer vision is the interpretation of video sequences. Traditionally, methods rely on grayscale or color data to infer semantic information.

In the past few years, new methods based on *Time-of-Flight* cameras have emerged. These cameras, hereinafter referred to as *ToF* (or *range*) cameras, produce low-resolution *range* images (also called *depth maps*), whose values indicate the distance between a pixel of the camera sensor and an object. Although distances measured by ToF cameras are relevant from a physical point of view, the technology has its own limitations: (1) since the size of a pixel on the sensor plane is larger than with a CCD camera, the resolution is relatively small, (2) distances are not measured precisely, (3) the calibration procedure is difficult, and (4) surface properties influence the reflections on objects, and consequently affect the measured distances.

To our knowledge, there is no theoretical model that embraces all the issues related to the acquisition of range data, but this hasn't stopped some companies to deliver products based on ToF cameras. Figure 1 shows a 3D model extracted from a range sensor; that model is used for an interactive game. For such an application, a complete 3D model of the scene cannot be deduced from the sole range image of a ToF camera. For example, there is no way to estimate the thickness of an object from a frontal depth map. Consequently, an

*has a grant funded by the FRIA, Belgium

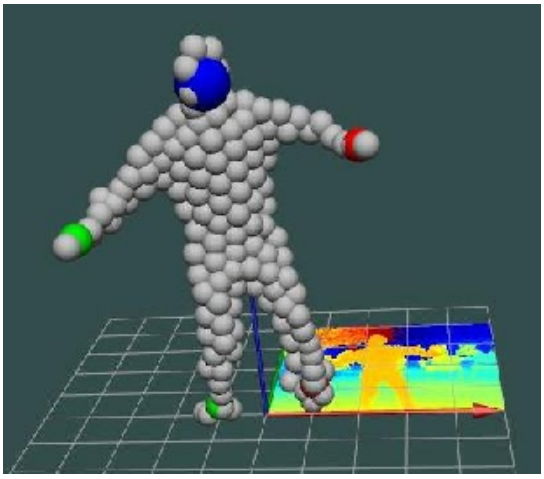

Figure 1: A 3D model reconstructed from a range image (image reproduced with the permission of SOFTKINETIC `http://www.softkinetic.net`).

elaborated model is required for an application that handles 3D objects in conjunction with ToF cameras.

In this paper, we detail a different approach that deals with pixels instead of objects. Such a pixel-based method allows to limit the number of assumptions about the scene and ignores the notion of any 3D model. More precisely, we aim to analyze the dynamic content of a scene by applying a background subtraction technique on a depth map. Our method is complementary to 3D model-based approaches as it can be used as a pre-processing step to locate objects in the foreground.

Background subtraction consists in separating pixels belonging to the background, where no motion is detected, from pixels of moving objects contained in the foreground. Silvestre [13] has briefly compared several background subtraction techniques on range maps for the purpose of video-surveillance. However, if one aims at an interactive application, it might not be sufficient to accurately segment the users, especially when they are located close to background objects. This is a similar problem to the confusion between background and foreground color occurring with color cameras: if a person's colors match those of the background, the person cannot be correctly detected.

In this paper, we propose to counter the aforementioned problems by combining depth and color information to enhance the robustness of a background subtraction algorithm.

We discuss the principles and limitations of Time-of-Flight cameras in Sec. 2 and 3. In Sec. 4, we explain how to combine motion detections coming from different modalities. Experimental results are given in Sec. 5. Section 6 concludes the paper.

2 Principles of Time-of-Flight Cameras

Time-of-Flight (ToF) cameras have already been described in several technical publications [13, 7, 8, 2] and, therefore, we limit our discussions to the basic principles of PMD (Photonic Mixer Device) ToF cameras.

PMD-cameras illuminate the whole scene with an infrared light ($\lambda = 870$ nm) whose envelope is modulated in amplitude: $s(t) = a + b\ \cos(\omega t)$ (where $a > b > 0$, t is the time, and ω relates to a modulation frequency of 20 MHz). Each pixel of the sensor receives the sum

RGB values | distance d | amplitude A | intensity I

Figure 2: A color image and the 3 channels provided by a PMD-camera.

of a time-delayed and attenuated signal reflected by the scene plus some additional ambient light in a small solid angle. It is assumed that the receiver is only sensitive to infrared. The received signal is thus $r(t) = k_a + k_b\ s(t - \Delta t) = a' + b'\ \cos(\omega(t - \Delta t))$.

In a PMD-camera, the device continuously multiplies the received signal $r(t)$ with 4 internal signals, given by $f_\theta(t) = a + b\ \cos(\omega t + \theta\frac{\pi}{2})$ with $\theta \in \{0, 1, 2, 3\}$, and computes their intercorrelations cor_θ. If the period of integration (shutter time) T is a multiple of $2\pi/\omega = 50\,\mathrm{ns}$, then

$$\mathrm{cor}_\theta = \frac{1}{T}\int_{<T>} f_\theta(t) r(t)\ dt = aa' + \frac{bb'}{2}\ \cos(\omega\Delta t + \theta\frac{\pi}{2})\ . \tag{1}$$

The distance d between the camera and the target is estimated by $c\ \Delta t/2$, where $c \simeq 3\ 10^8$ m/s is the speed of light. It is computed using the four intercorrelation values. Note that there is a distance ambiguity after $\pi c/\omega \simeq 7.5$ m, due to the Δt-periodicity of cor_θ. If j represents the complex number, we have the following estimation for the distance

$$d = c\ \frac{\arg\left(\mathrm{cor}_0 - j\ \mathrm{cor}_1 - \mathrm{cor}_2 + j\ \mathrm{cor}_3\right)}{2\,\omega}\ . \tag{2}$$

The amplitude b' of the received signal is also provided. It is related to the peak-to-peak amplitude A of the intercorrelations:

$$A = bb' = \sqrt{\left(\mathrm{cor}_0 - \mathrm{cor}_2\right)^2 + \left(\mathrm{cor}_1 - \mathrm{cor}_3\right)^2}\ . \tag{3}$$

It measures the strength of the incoming signal. The amplitude obviously decreases as the distance between the sensor and the object increases. This has led some authors [9] to establish a relationship between A and a grayscale (luminance) image. But this relationship is incorrect in some cases like, for example, clouds seen through a window, where the amplitude A will be equal to 0 to the contrary of the luminance.

The continuous component a' of the received signal is the third information a PMD-camera can provide. It is expressed by the intensity I, and estimated as

$$I = aa' = \frac{\mathrm{cor}_0 + \mathrm{cor}_2}{2} = \frac{\mathrm{cor}_1 + \mathrm{cor}_3}{2}\ . \tag{4}$$

In conclusion, as shown in Fig. 2, a PMD camera provides three values per pixel: d, A and I. These values can be interpreted as follows:

- d is the estimated distance between the illuminated object and the sensor,
- A estimates the quality of the signal used for the determination of d, and
- I is related to the temporal average amount of received infrared light.

One must note that the aforementioned interpretations must be handled with care: the intuitive interpretation of the three channels provided by a range camera is, at best, delicate. In the next section, we detail the various limitations that must be taken into account when using a PMD range camera.

3 Limitations of Time-of-Flight Cameras

The signals given by a PMD-camera (d, A and I) are imperfect. Differences exist between the theoretical principle and its implementation: some delays are introduced by the electronics and the wave envelope isn't a perfect sinusoid [8]. Furthermore, a static error, called *wiggling effect*, has been observed on d. It is an oscillating and periodic function of the true distance [7, 8]. As noted in [4], a temporal analysis of d shows that its standard deviation is proportional to its mean. Thus, the variance of the noise on d is depth-dependent. [12] showed that both d and A depend on the shutter time T. As of today, there is no available theoretical model explaining all these effects.

Other imperfections come from the scene itself [7]. Because of the low resolution of the sensor, each pixel corresponds to a wide solid angle, leading to errors. Furthermore, the estimated distance depends on the reflectivity and the orientation of the observed object [7]. Moreover, as Fig. 2 shows, artefacts appear in the A and I channels near distance discontinuities. Finally, our experiments showed a dependence of A and I over multi-paths.

Some authors tried to reduce the error on the estimated distances using the information contained in the A or I channels [7, 9]. But due to error dependencies, A, I and T should be used simultaneously to correct d. Even then, getting a perfect measure is impossible if the content of the scene is unknown.

4 Combining Depth with Other Modalities

The three channels provided by the PMD-camera are not suited for precise distance measurements, but it is possible to simplify the problem. We model all the defects on the channels as an additive noise and try to recover useful information from these noisy signals using a widely used video segmentation technique: background subtraction, which is described below.

4.1 Motion Detection and Background Subtraction

Background subtraction is one of the most ubiquitous automatic video content analysis technique. Its purpose is to isolate the parts of the scene corresponding to moving objects. For an interactive application, the point of using background subtraction is straightforward: the moving objects detected by the algorithm correspond either to the users or to physical objects they interact with.

Numerous methods for background subtraction have been proposed over the past years (see [10, 11] for surveys). From a review of the literature on the subject, it appears that recent *sample-based* techniques [1, 3, 14] are particularly well-suited for our needs: they are fast, versatile, and resilient to important amounts of noise. Furthermore, these techniques are pixel-based: they analyze the temporal evolution of each pixel independently. As a result, they do not make any assumption about the content of the images they process.

We use the ViBe algorithm presented in [1]. Our main motivation is the robustness to noise exhibited by ViBe as shown in the experiments of [1]. ViBe has other advantages: it is fast, simple to implement and shows a great accuracy on the contours of the silhouettes.

Hereinafter we explain our motivations to apply ViBe on conventional color or grayscale images, and on range images separately. Then in Sec. 4.2, we combine both techniques.

4.1.1 Application to a Color or Grayscale Image.

For most interactive applications, a background subtraction technique is a useful tool. The sole binary silhouettes of the users can provide sufficient amount of information in simple applications who often take advantage of the high precision of the silhouettes obtained with conventional cameras.

However, the use of RGB or grayscale images has a few intrinsic limitations. Due to the lack of 3D information, the system cannot recognize simple actions such as pointing a finger to an area of the screen. Furthermore, background subtraction itself imposes major restrictions on operating conditions: the illumination of the scene must be controlled and the colors of the users may not match those of the background.

4.1.2 Application to a Range Image.

If applied on a range image, background subtraction does not suffer from the two limitations mentioned in the above paragraph. Indeed, the range camera uses its own light source and is only slightly affected by the ambient lighting. It can even be used in complete darkness. Furthermore, since it does not use color information, it is not sensitive to users' colors. Unfortunately, a problem occurs when the users are physically too close to the background of the scene. In such a worst-case situation, it is impossible to discriminate between the objects and the background because of the noisy nature of depth maps. Furthermore, due to the low resolution of the PMD sensor, the use of ViBe on the sole depth map cannot produce precise segmentation.

From the above discussions, it appears that an optimal solution to get robust silhouettes consists in combining the benefits of several motion segmentations obtained from various modalities. In the next section, we present a method to combine both segmentation maps and describe the technical issues raised by such a combination.

4.2 Combining Color, Depth, and Motion

Our experimental setup is made of an RGB-camera and a PMD-camera fixed on a common support, one on top of the other. Both cameras are equipped with similar objectives, but their field of view are different. The first major issue in using a couple of cameras is that of image registration.

4.2.1 Image Registration.

With a precise distance channel, we could theoretically link the two focal planes. First, a projective model must be chosen for each camera and internal and external parameters must be computed. The distance estimation then helps to locate which 3D point is projected on the PMD-pixel. This point can be reprojected with the model of the RGB-camera to get the corresponding RGB-pixel. This process is valid as long as no other object stands between the 3D point and the RGB-camera. This correspondence has to be computed for each PMD-pixel.

The two cameras can follow the pin-hole model [5]. Their sensor geometries are similar and their optical systems are identical. The determination of the internal and external parameters of the RGB-camera is a classical problem [6, 15]. The calibration of a PMD-camera is more difficult. First, its low resolution (160×120 pixels) makes it difficult to associate a pixel with a 3D point. Unfortunately, there is no way to determine the external parameters without using

Table 1: Usability of motion detection on the grayscale channel of the color camera and on each channel of the range camera.
[1]When the difference in reflectance between the target and the background is small, amplitude and intensity cannot be used if the distance between them is short.
[2]Amplitude and intensity can be used if the difference in reflectance between the target and the background is large.

Operating conditions	grayscale	distance	amplitude	intensity
low target/background contrast	no	yes	yes/no[1]	yes/no[1]
small target to background distance	yes	no	yes/no[2]	yes/no[2]
distance to background larger than 7.5 m	yes	no	yes	yes
low scene illumination	no	yes	yes	yes
fluctuations in scene illumination	no	yes	yes	yes

these correspondences. Second, it is hard to build an appropriate calibration object. A plate with holes is inadequate because there are significant artefacts near the edges. Moreover, many paints don't absorb infrared light (black paintings are not necessarily the best ones). This complicates the determination of the internal parameters.

An alternative to the calibration of both cameras is to establish a static mapping between PMD- and RGB-pixels. Unfortunately, the linearity and the continuity of this mapping are only guaranteed if the observed world is flat or if both cameras share the same optical center.

In our application, a very precise matching is not required and we deliberately want to avoid both an uncertain calibration and any assumptions about to the scene content. This led us to use a static affine mapping between the RGB image and the depth map. From our experiments, it has proven to be sufficient, as the optical centers of both cameras are close to each other and the targets are far from the camera.

4.2.2 Motion Segmentations Combination.

We consider the behavior of the background extraction algorithm on the grayscale channel of a color camera and on the different channels of a range camera. An extended case study of the usability of the 4 channels is given in Table 1.

We detect motion on the grayscale image and on each channel of the range camera. Combining the motion maps allows us to deal with most of the practical scenarios. Three practical examples of successful combination follow:

1. If the target is physically close the background and the scene illumination is low, distance and grayscale information are useless for motion segmentation. However, amplitude and intensity channels are still usable if the reflectance properties of the background and the target differ.

2. Under low illumination conditions or when the targets and the background looks similar, depth is a meaningful channel for motion detection.

3. Background subtraction on the luminance channel is impossible in presence of large fluctuations of the scene illumination, but the channels of a PMD-camera are not perturbed by this phenomenon.

The complete segmentation process is drawn in Fig. 3. To refine the segmentation maps near the edges, the motion masks of I and A are eroded. An affine transform is used to map the images generated from the two cameras. A logical (non-exclusive) "or" is used to combine the different foregrounds and noise is removed by morphological filtering.

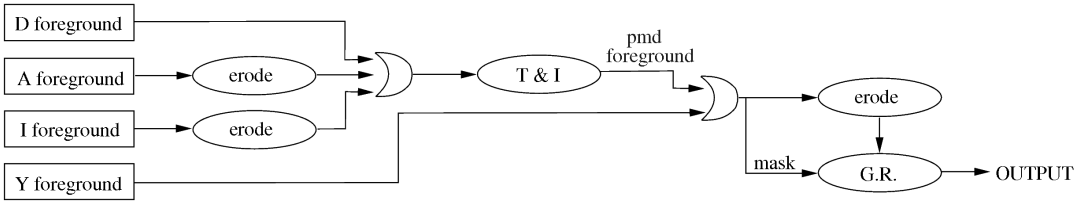

Figure 3: Complete description of our method. "*G.R.*" stands for "*Geodesic Reconstruction*" and "*T&I*" stands for "*Transformation and Interpolation*".

5 Results and Application

This section presents some segmentation results. We also illustrate our method by projecting the segmented depth map and RGB texture in a 3D engine.

5.1 Motion Segmentation Results

As shown on Fig. 4, users too close to the background cannot be correctly segmented given the sole depth map. As a matter of fact, the minimal distance at which the detection becomes possible is conditioned by the amount of noise present in the depth channel. However, by taking advantage of the grayscale image, our algorithm manages to successfully segment those users.

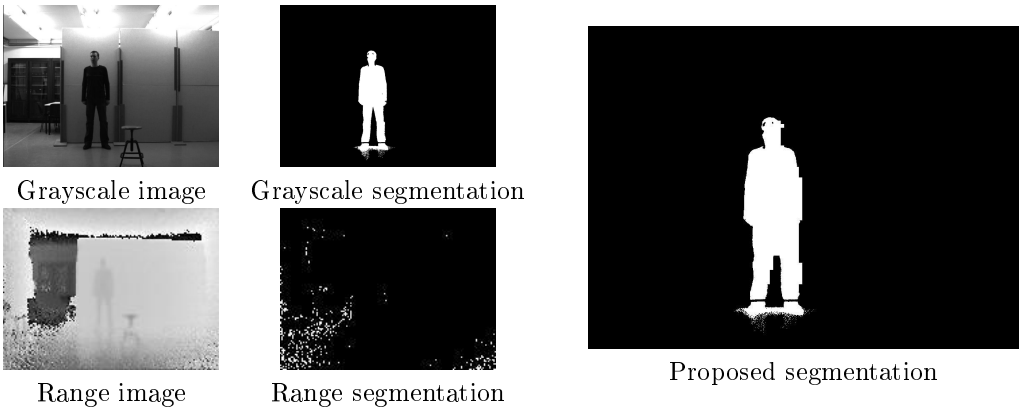

Figure 4: This figure shows that the failure of the background subtraction algorithm in one of the used modalities (*range* in this case) does not harm the motion detection resulting from the proposed fusion method.

Figure 5 illustrates a case with poor motion detections for all the modalities. Since most of the locations of the segmentation errors differ for each modality, the proposed method showed to be able to produce accurate results, even in such a pathological situation.

5.2 Example of an Immersive Application

Finally we combine in real time the depth signal with the resulting segmentation map to project users in a virtual 3D environment. A 3D mesh is constructed on the basis of the segmented depth signal. By using the affine transform described previously to map the RGB

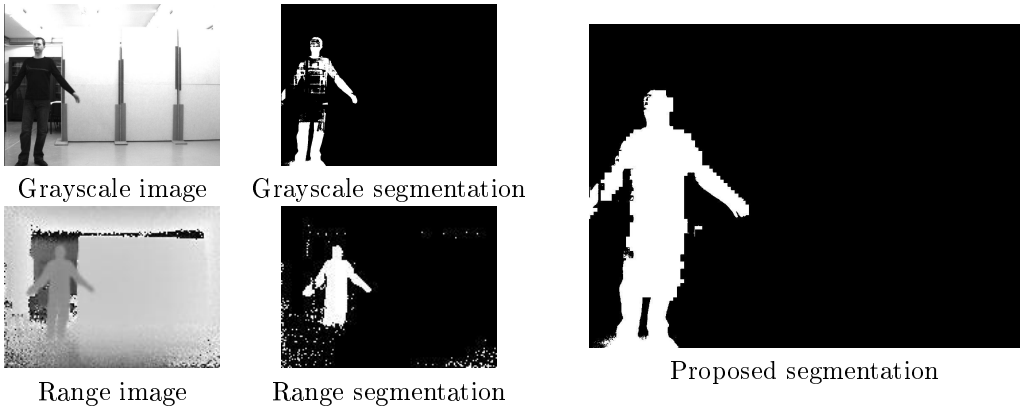

Figure 5: Even when none of the used modalities produces a satisfactory motion detection, the proposed fusion method is still able to successfully segment the video stream.

image onto the 3D mesh, we create a valid 3D representation of the scene. As show in Fig. 6, we achieve a very convincing 3D representation of the scene without any 3D model, despite the pixel-based nature of our approach. It is worth mentioning that the whole process runs in real time.

Figure 6: Application of the proposed algorithm in an immersive application. The depth signal serves to construct a mesh and an RGB texture is mapped on that mesh.

6 Conclusions

This paper presents a novel approach for the interpretation of a video scene which takes advantages of the signals provided by a PMD Time-of-Flight camera. By combining these signals with those of a RGB camera and processing them with a background subtraction algorithm, we are able to extract meaningful information from the depth signals, despite their noisy nature. As a showcase for a practical scenario, we show how the RGB image, the depth map and the motion segmentation can be combined for an interactive application or an immersive human-machine interface.

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)*, pages 945–948, April 2009.

[2] Francois Blais. Review of 20 years of range sensor development. *Journal of Electronic Imaging*, 13(1):231–243, 2004.

[3] A. Elgammal, D. Harwood, and L. Davis. Non-parametric model for background subtraction. In *ECCV '00: Proceedings of the 6th European Conference on Computer Vision-Part II*, pages 751–767, London, UK, 2000. Springer-Verlag.

[4] D. Falie and V. Buzuloiu. Noise characteristics of 3D time-of-flight cameras. In *International Symposium on Signals, Circuits and Systems (ISSCS)*, volume 1, pages 1–4, July 2007.

[5] S. Fuchs and S. May. Calibration and registration for precise surface reconstruction with ToF cameras. In *DAGM Dyn3D Workshop*, September 2007.

[6] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[7] M. Lindner and A. Kolb. Lateral and depth calibration of PMD-distance sensors. In *Advances in Visual Computing*, volume 2, pages 524–533. Springer, 2006.

[8] M. Lindner and A. Kolb. Calibration of the intensity-related distance error of the PMD TOF-camera. In *SPIE: Intelligent Robots and Computer Vision XXV*, volume 6764, pages 6764–35, 2007.

[9] S. Oprisescu, D. Falie, M. Ciuc, and V. Buzuloiu. Measurements with ToF cameras and their necessary corrections. In *International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4, July 2007.

[10] M. Piccardi. Background subtraction techniques: a review. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3099–3104, 2004.

[11] R. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image change detection algorithms: A systematic survey. *IEEE transactions on Image Processing*, 14(3):294–307, 3 2005.

[12] J. Radmer, P. Fusté, H. Schmidt, and J. Krüger. Incident light related distance error study and calibration of the PMD-range imaging camera. In IEEE Computer Society, editor, *Conference on Computer Vision and Pattern Recognition*, pages 23–28, Piscataway, NJ, 2008.

[13] D. Silvestre. Video surveillance using a time-of-flight camera. Master's thesis, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 2007.

[14] H. Wang and D. Suter. A consensus-based method for tracking: Modelling background scenario and foreground appearance. *Pattern Recognition*, 40(3):1091–1105, 2007.

[15] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.