REAL-TIME PROCESSING OF DEPTH AND COLOR VIDEO STREAMS TO IMPROVE THE RELIABILITY OF DEPTH MAPS

Sébastien Piérard , Jérôme Leens, Marc Van Droogenbroeck

INTELSIG Laboratory, Montefiore Institute, University of Liège, Belgium

ABSTRACT

Depth is a useful information in vision to understand the geometrical properties of an environment. Depth is traditionally computed in terms of a disparity map acquired by a stereoscopic system but, over the last few years, several manufacturers have released single-lens cameras that directly capture depth information (also called *range*). This is an important technological breakthrough although range signals remain difficult to handle in practice, due to many reasons (low resolution, noise, low framerate, ...). Practitioners still struggle to use range data in their applications. The purpose of this paper is to give a brief introduction to range data (captured with a camera), discuss common limitations, and propose techniques to cope with difficulties typically encountered with range cameras. These techniques are based on a simultaneous view of the scene by a color and a depth camera that are combined to improve their interpretation in real time.

Index Terms— Range camera, depth, motion detection

1. INTRODUCTION

One of the main tasks in computer vision is the interpretation of video sequences. Classical methods rely on grayscale or color data to infer semantic information. But cameras that measure distances (called 3D or depth cameras in this paper) pave the way to new techniques in computer vision. There are three main application areas:

- Illumination-independent applications.
  As described hereafter, depth cameras use their own (invisible) light. They are thus well suited for interactive applications in environments such as projection rooms. In such applications, color cameras are inefficient and infrared cameras fail to discriminate between "cold" objects.
- 3D-driven applications.
  Depth cameras are also interesting in applications related to 3D vision. With a color camera, even a simple action, such as pointing a finger to a screen, is a difficult movement to recognize, because color cameras capture color and shape information but no 3D information. Stereoscopic vision is an alternative to retrieve range information but, in an uncontrolled environment, the potential lack of texture makes the system to fail. It is not surprising that many companies involved in interactive gaming have expressed their interest for 3D data. Among them, *Microsoft* has unveiled a new interface for the *Xbox* 360 based on a low cost depth camera. The use of such cameras in immersion games does not require any controller (such as the *Wiimote* of *Nintendo*), in contrast to current products. Some people even claim that depth cameras could rapidly become common in consumer products.

S. Piérard has a grant funded by the FRIA, Belgium.

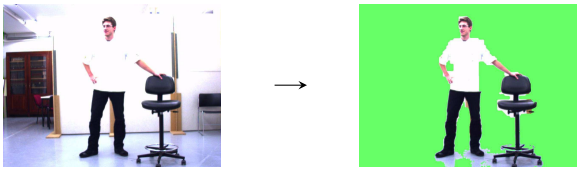

Fig. 1. The goal of foreground segmentation is to isolate the users and objects located in the foreground.

- Applications needing a foreground/background discrimination.
  Depth cameras and color cameras are complementary to isolate the foreground from the background (see Fig 1). Indeed, a foreground/background segmentation based on colors is bound to fail if the objects in the foreground have the same color as the background. However situations where users and objects have the same color and depth as the background remain exceptional. A combination of color and depth will prove itself useful for applications that do not require precise 3D information, such as video-surveillance.

There are also application areas where 3D cameras have not been used because of their current shortcomings, like metrology or the movie industry. For such areas, the signals produced by a 3D camera have to be improved or used in conjunction with other modalities (laser measures, color images, etc).

The remainder of this paper is organized as follows. Section 2 explains the principles of 3D acquisition by time-of-flight cameras. Then, Section 3 discusses how depth signals can be used in practice. We summarize some possible improvements in Section 4, and conclude in Section 5.

2. TIME-OF-FLIGHT CAMERAS

The basic principle of depth cameras is to measure distances between the camera and visible points of the scene. As the direct measure of the distance is intractable, one uses a measure of time instead to derive the distance. A depth camera emits a signal that is reflected on the scene, and measures the time Δt needed by the signal to go and return. This explains why depth cameras are sometimes called time-of-flight (ToF) cameras. If the emitter and the receiver are punctual and located at the same place, then the distance is $d = c\Delta t/2$, where c is the speed of the signal ($c \simeq 3 \times 10^8$ m/s for light).

2.1. PMD cameras

Among ToF technologies, PMD (Photonic Mixer Device) cameras are widespread. These cameras have already been described in several technical publications [2, 5, 6, 8]; therefore, we limit our dis-

Fig. 2. A PMD camera (*PMD[vision]19k*). Source lights (infrared LEDs) are located on both sides of the sensor.

cussions to the basic principles. PMD cameras illuminate the whole scene with invisible light modulated in amplitude. The modulating signal is chosen to be

$$s(t) = a + b\cos(\omega t),$$

where $a > b > 0$, t is the time, $\omega = 2\pi f$, and f is the modulating frequency (not to be confused with the frequency of the carrier). Usually, infrared light is used as the carrier (at a 870 nm wavelength), and the modulating frequency f is 20 MHz. Pixel sensors receive a time-delayed and attenuated version of the source signal, plus some ambient light:

$$r(t) = a' + b'\cos(\omega(t - \Delta t)).$$

The challenge consists in an appropriate interpretation of the three parameters a', b', and Δt for each pixel, especially for deriving the depth map. Note that a, b, and more importantly the original phase ωt, are known to the receiver.

2.2. The PMD signals

The PMD sensors provide three channels per pixel (or two for outdoor cameras). The channels are shown in Figure 3:

1. the *distance* d is proportional to the phase shift (and thus to Δt) between the emitted and the received signals. 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 b' 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. the *intensity* I is proportional to the direct component a' of the received signal, and relates to the global luminance of the scene. Outdoor PMD cameras do not provide this third channel, because their electronic circuits are designed to avoid saturation.

The intuitive interpretation of the three channels provided by a range camera is, at best, delicate.

2.3. The PMD operations and limitations

The task of a PMD device is to derive a', b', and Δt from the received signal $r(t)$. To achieve this, the received signal $r(t)$ is multiplied with 4 signals (internally to the device), expressed by

$$f_\theta(t) = a + b\cos\left(\omega t + \theta\frac{\pi}{2}\right) \quad \text{with } \theta \in \{0, 1, 2, 3\}\,.$$

Then the device computes four intercorrelation signals, cor_θ, whose integration period (shutter time) T is chosen to be a multiple of

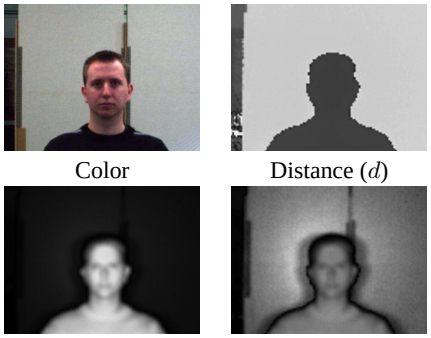

Fig. 3. A color image and the 3 PMD channels.

$2\pi/\omega = 50$ ns, so that cor_θ does not depend on time. It can be shown that

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

and that cor_0, cor_1, cor_2, and cor_3 suffice to derive the three channels d, A, and I (see [4] for more details). For example, the distance is estimated as

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

where j denotes the complex number and $\arg$ is the function that gives the argument of a complex number. This expression is theoretically correct, but the quality of the measures is subject to caution because we still lack a theoretical model explaining the exact behavior of the sensor. In addition, note that the ambient infrared light is required to be statistically stationary to have cor_θ values being independent from time. It is otherwise impossible to derive the values of d, A, and I.

Despite its attractiveness, the PMD technology has its own limitations:

- Many sources of noise disturb the received signal. As d, A, I are evaluated indirectly (via four intercorrelations), the estimators mix several kinds of noise. In particular, d is not an accurate distance estimation. Increasing the shutter time T reduces the variance, but at the price of increasing the bias. Depth calibration of PMD sensors is an active field of research aiming at canceling this bias. Not surprisingly, it has been shown that the bias is related to both the distance and the amount of received infrared light [7]. Moreover, taking a larger T implies that the minimal distance between the camera and objects has to be increased to avoid saturation. Measuring depth with PMD cameras is thus not straightforward.
- The two A and I channels measure the amount of received infrared light. Thus, they depend on both the nominal distance between the camera and objects (the power attenuation increases with the distance), and the properties of objects (absorption/reflection coefficients and orientation).
- The current PMD cameras have a fairly small number of pixels (about 200×200). This limited resolution introduces practical limitations to calibration. While the optical calibration (*intrinsic* parameters) is still possible with a well-known calibration object, the spatial calibration (*extrinsic* parameters) is almost impossible. Indeed, spatial calibration requires to

register 3D points with pixels. However, due to a low resolution, each pixel corresponds to a wide solid angle, leading to inaccuracies that are unacceptable in many applications.

- It is impossible to use multiple PMD cameras in the same room because the carrier frequency is unique, even with different modulating frequencies.
- One can show that $\text{cor}_\theta(\Delta t) = \text{cor}_\theta(\Delta t + 1/f)$. Distance computations based only on the phase delay are therefore ambiguous up to a $c\frac{1}{2f}$ factor, equal to 7.5 m if $f = 20$ MHz. For example, an object located at 9 m is considered to be at a distance of $9 - 7.5 = 1.5$ m.
- Because of persistence effects, when a fast movement occurs, a trail is observed in the three channels of the PMD camera. If the observed person or object moves quickly, a ghost appear in the images.
- The position of the light source influences the quality of the measured distances. As the light sources of PMD cameras are not located at the optical center of the sensor, shadows are observed in the A and I channels. When the light sources are laterally positioned (as in Fig. 2), shadows are always on the left and on the right of the foreground objects.

3. USING THE DEPTH SIGNALS

The channels provided by PMD cameras are not suited for precise distance measurements. It appears that a depth calibration is not possible if one wants to deal with unknown objects at unknown distances. It is the relationship between the PMD channels that hinders the calibration process: d and A depend on each other.

However, PMD cameras are useful because it is possible to detect temporal modifications on the channels, even in the presence of additive noise. This is performed by *background subtraction* (also called *foreground detection*) algorithms, which are one of the most ubiquitous automatic video content analysis technique. Its goal is to isolate moving people and objects in the scene.

Apparently, it should be sufficient to apply a threshold on the distance channel to perform a foreground detection. In practice however, thresholding the d channel does not work well for many reasons, like the presence of noise, saturation effects, distance ambiguity, etc.

From our analysis of background subtraction algorithms, it appears that recent *sample-based* techniques [1, 3] are particularly well-suited to fill our needs: they are fast (real-time), versatile, and resilient to important amounts of noise. In addition, these techniques are pixel-based, which means that each pixel is processed independently. This is an important advantage because it avoids that noise is spread over neighboring pixels during the segmentation process. Moreover there is no need to have a precise statistical model for pixel values so that they can accommodate to any type of image channel.

Background subtraction has been intensively developed for grayscale or color cameras but, in our case, we apply it on the PMD channels. The result is that the background subtraction is performed on channels that relate to the physical distance. Consequently, a PMD foreground detection not only detects moving objects but also evaluates the distance between the objects and the camera. It can, for example, distinguish between moving objects in the close foreground and moving objects in the background.

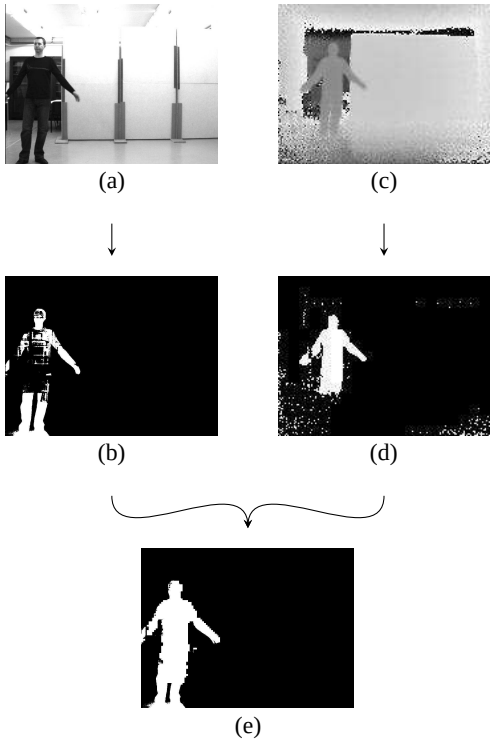

Fig. 4. Combination of background subtractions on a grayscale image and on the depth map. (a) grayscale image, (b) foreground of (a), (c) distance map, (d) segmentation of (c), and (e) final segmentation.

4. PROPOSED ENHANCEMENTS

Many applications based on PMD cameras require an adequate foreground/background separation. However, the foreground segmentation is imperfect because of the PMD shortcomings (see Section 2.3). Therefore it is compulsory to improve the foreground detection. Hereinafter, three enhancement methods are suggested.

4.1. Improved foreground/background discrimination via motion detection and combination of color and depth data

As explained in the introduction, range data and color data are complementary to isolate the foreground from the background. Therefore, we combine the information provided by a color camera and a depth camera. This permits to detect all the moving objects despite the uncertainty on distance or color confusion.

We apply a background subtraction algorithm on the three PMD channels separately and merge the results to obtain a PMD foreground. This foreground is then merged with the one of the color camera to build the final segmentation (see Fig. 4 for an illustration).

4.2. Suppression of lateral shadowing effects in the PMD foreground

Infrared sources located on the left and on the right of the sensor project shadows on the right and left sides, respectively, of the foreground. The shadows affect both channels A and I, but not d (to a first approximation). Because we use the three PMD channels to

(a) (b) (c)

Fig. 5. Removal of lateral shadows in the foreground. (a) intensity channel I with shadows, (b) distance channel d. (c) PMD foregrounds prior and after correction. The shadows (in red) are significant.

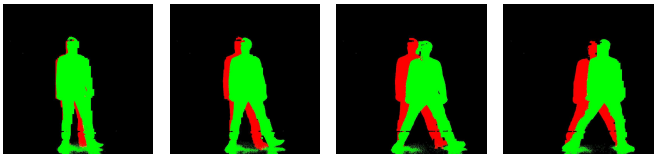

Fig. 6. Illustration and correction of the persistence effects. The areas of persistence (in red) are significant.

build the PMD foreground, this one is also affected by the shadows. However, it is possible to rectify the PMD foreground by searching for strong transitions in d. Shadows shift the lateral border only for large distances between the foreground and background. Thus, for strong transitions on d inside the PMD foreground, we displace its contour towards the center. Fig. 5 shows the result of this correction method.

4.3. Removal of persistence effects

Our experiments show that the duration of the persistence effects on the PMD modality is limited to only one frame. The key of our solution consists in comparing the foregrounds computed on the PMD and color cameras. Persistence occurs in areas of pixels that have left the color background and are still present in the PMD foreground. These pixel areas are to be removed from the global segmentation map. Fig. 6 illustrates the regions affected by persistence (in red).

5. CONCLUSIONS

As any other technology, the PMD technology has its own limitations. This paper provides an overview of the principles of operation of this technology and present some solutions to enhance the interpretation of the distance measurements. It appears that 3D cameras are an excellent tool for interpretation as long as applications concentrate on the foreground.

6. 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] F. 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 *Proceedings of the 6th European Conference on Computer Vision-Part II*, pages 751–767, London, UK, 2000. Springer-Verlag.

[4] 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*, volume 5815 of *Lecture Notes in Computer Science*, pages 104–113. Springer, 2009.

[5] 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.

[6] 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.

[7] 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 International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23–28, Piscataway, NJ, 2008.

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