FAST REAL-TIME SEGMENTATION AND TRACKING OF MULTIPLE SUBJECTS BY TIME-OF-FLIGHT CAMERA

A New Approach for Real-time Multimedia Applications with 3D Camera Sensor

Piercarlo Dondi and Luca Lombardi
Department of Computer Engineering and System Science, University of Pavia, Via Ferrata 1, Pavia, Italy

Keywords: Time-of-Flight camera, Segmentation, Tracking, Mixed reality.

Abstract: Time-of-Flight cameras are a new kind of sensors that use near-infrared light to provide distance measures of an environment. In this paper we present a very fast method for real-time segmentation and tracking, that exploits the peculiar characteristics of these devices. The foreground segmentation is achieved by a dynamic thresholding and region growing: an appropriate correction based on flexible intensity thresholding and mathematical morphology is used to partially compensate one of the most common problem of the TOF cameras, the noise generated by sun light. By the use of a Kalman filter for tracking the retrieved objects the system is able to correctly handle the occlusions and to follow multiple objects placed at different distances. The proposed system is our basic step for complex multimedia applications, such as augmented reality. An example of mixed reality that includes the integration of color information, supplied by a webcam is shown in the experimental results.

1 INTRODUCTION

The introduction of Time-of-Flight (TOF) cameras (Oggier et al., 2004) made possible real-time depth measurement using a single compact sensor in spite of previous multi-camera systems, like stereo cams.

The researches of the last few years showed a large interest on this kind of devices in many fields related to computer vision and computer graphics, like 3D modeling, scene recostruction, user interaction or segmentation and tracking of moving people (Kolb et al., 2010). In our research we are focused primarily on these last two topics. Our main purpose is the development of a fast and accurate system for segmentation and tracking of multiple subjects, than can be used as basic step for multimedia applications, as augmented reality or human-machine interaction.

For achieving this goal we have considered strengths and weaknesses of these new kind of devices. So we have designed a real-time foreground segmentation algorithm that exploits the peculiar data provided by TOF cameras and also compensates one of this most recurrent problem, the noise introduced by sun light.

Through the successive tracking phase based on Kalman filter the system is able to follow multiple subjects also in presence of occlusions and guarantees the concurrent retrieval of subjects placed at different distances.

The paper is organized as follow: section 2 provides an overview of TOF cameras; section 3 presents the state of art of TOF based segmentation algorithms and describes our solution with the proposed method for compensate the sunlight interferences; section 4 analyzes our tracking method and its integration with segmentation; in section 5 we show the experimental results and in section 6 our conclusions.

2 TIME-OF-FLIGHT CAMERA

Time-of-Flight cameras are active imaging sensors that use laser light in near infrared spectrum to provide distance measures of the scene. There are two main solutions at the base of these devices: pulsed light and modulated light. In the first case a coherent wavefront hits the target and high frequency photon gating measures the return time-of-flight; in the second, the emitted light is modulated and time-of-flight is measured by phase delay detection.

TOF cameras supply some advantages respect to laser scanners or stereo cameras: they do not use any moving mechanical component, can work at real-time

Dondi P. and Lombardi L..
582 FAST REAL-TIME SEGMENTATION AND TRACKING OF MULTIPLE SUBJECTS BY TIME-OF-FLIGHT CAMERA - A New Approach for Real-time Multimedia Applications with 3D Camera Sensor.
DOI: 10.5220/0003322505820587
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2011), pages 582-587
ISBN: 978-989-8425-47-8

(18-20 fps), are insensible to shadows and can measure 3D distance in any kind of scenario. Artificial illumination sources do not interact with the sensor, but the presence of sun light introduces a significant noise. A TOF camera has a nominal range of about 10 m: noise caused by scattering, multi-paths and environment light can reduce this value, so the useful range is generally between 2 and 5 meters (Oprisescu et al., 2007).

In our project we utilize the SR3000 realized by MESA Imaging (Oggier et al., 2004), a modulated-light camera that we use at 20MHz. Its active sources emit in the near infrared around 850nm. The SR3000 supplies two maps per frame with a resolution of 176x144 pixels: the former contains distance information and the other one measures the intensity of light reflected by objects. The sensor is completely indifferent to visible light, so values of intensity depend only by light in near infrared. As a consequence closed objects appear more clear because reflect more light, instead faraway objects result to be darker.

3 SEGMENTATION

There are some different approachs for TOF-based segmentation. An interesting solution, based only on the depth data, is proposed in (Parvizi and Wu, 2008). The extraction and the association of the objects are made analyzing the probability density function of the depth and the distribution of its histogram.

Moreover the most common method is certainly the background subtraction, a classical computer vision tecnique, widely used for video surveillance and for tracking applications. This method is most often pixel-based and only rarely region-based; newly acquired frames are compared to the model and pixels which differ significantly are selected as foreground. In some works the background model is estimated using only the data providing by TOF-camera (Hansen et al., 2008) or (Bevilacqua et al., 2006), while in other cases is made using a combination of multiple cameras (Guomundsson et al., 2008). In particular the integration of color and depth data for background modeling is largely used for multimedia applications, like in (Crabb et al., 2008) for background substitution, or in (Bartczak et al., 2008) for create a 3D ambient for mixed reality.

However, background subtraction suffers from known problems like ghosts appearing when background objects changes or absorption of immobile persons. Background model generation can also be computationally expensive, expecially if it needs a high resolution 3D model.

Also other alternative approaches take advantage of the combination of traditional and TOF cameras. In (Santrac et al., 2006) the depth information are used for select the best input area for a color based segmentation algorithm (SIOX). While in (Bleiweiss and Werman, 2009) a fusion of colors and depth data is employed for creating a new segmentation and tracking method. This solution, based on mean shift algorithm, is intended to compensate the respective weaknesses of the two different kind of sensors.

The presented segmentation method is designed so as not to need any preprocessing operations or a priori knowledge of the environment or of the objects. It can be subdivided in two main phases: a first thresholding of the distance map based on the correspondent values of intensity image, followed by a region growing that starts from seeds planted on peaks of the intensity map.

3.1 Thresholding and Region Growing

Considering the characteristic of the TOF camera summarized in section 2, we decided to use the data of intensity map as a guide to restrict the area of investigation in the range map and to find good candidates for becoming seeds.

We estimate an opportune intensity threshold (λ_{seed}) applying the Otsu's method. This parameter is used to define the set of seeds S (see formula 1).

$$\{I_x > \lambda_{seed}, \|P_x - P_s\| > \gamma, \gamma > 1\} \rightarrow \{P_x \in S\} \quad (1)$$

P_x is a point of the distance map, I_x is its corresponded intensity value and P_s is the last seed found. The presence of a control about the distance between seeds guarantees a better distribution of them and reduces significantly their number in order to decrease the time needed for the following growing step.

The similarity measure S between a cluster pixel x and a neighboring pixel y is defined in 2:

$$S(x,y) = |\mu_x - D_y| \quad (2)$$

D_y is the distance value of pixel y and μ_x is a local parameter related to the mean distance value around x (see equation 5). The lower is S, the more similar are the pixels. When a seed is planted, μ_x is initialized to D_x. Considering a 4-connected neighborhood, a pixel x belonging to a cluster C absorbs a neighbor y according to the following conditions:

$$\{x \in C, S(x,y) < \theta, I_y > \lambda\} \rightarrow \{y \in C\} \quad (3)$$

λ is an intensity threshold proportional to λ_{seed} dynamically calculate for every frame using equation 4.

$$\lambda = k * \lambda_{seed}, k \in [0.25; 0.33] \quad (4)$$

θ is a constant parameter experimentally estimated. Our tests with multiple sequences of data estabilsh that θ must assume a value power of 2 major of 512 for maintain a good clusters separation. An optimal default choice is 1024.

When a neighbor y of seed x is absorbed, we compute the average distance value μ_y in an incremental manner as follows:

$$\mu_y = \frac{\mu_x * \alpha + D_y}{\alpha + 1} \quad (5)$$

Parameter α is a learning factor of the local mean of D. If pixel y has exactly α neighbors in the cluster, and if the mean of D in this neighbor is exactly μ_x, then μ_y becomes the mean of D when y is added to the cluster.

Every region grows excluding the just analyzed pixels from successive steps. The process is iterated for all seeds in order of descending intensity. Regions too small, with dimension inferior to a fixed value, are discarded.

Our approach is faster than methods that use global region statistics, like e.g. centroid region growing, where the order in which boundary pixels are tested is significant. μ_y depends only on the history of pixel absorptions until y is first reached by a growing front, and not from later steps. Thus, as soon as a pixel y is reached by the cluster boundary, it can be tested for absorption.

The locality of our approach tolerates greater variations of map values inside a cluster because it produces transitive closures of the similarity S. Both head and shoulders of the same person, lying at slightly different distance from the camera, are more likely to be segmented as the same cluster, rather than two different clusters (Bianchi et al., 2009).

The advantages of the proposed region growing can be summarized in good quality of boundaries (intrinsic noise rejection), independence of background models, and independence of shape models.

3.2 Improvements for noised Conditions

The proposed approach is very fast and ensures a good compromise between the computational time and the precision of the results. A previous analysis (Bianchi et al., 2009) has described its behavior in term of correctness (the percentage of correctly extracted foreground data) and completeness (the percentage of the reference data that is explained by the extracted data) in funciton of the parameter λ.

In optimal conditions, with no noise generated by sun light, the tests have scored between 94% and 97% in correctness and between 92% and 96% in completeness (see red dashed line and blue continuous line in figure 1). Otherwise in a more general situation, like a room with a window, the system presents a predictable reduction of completeness (82%-85%), also maintaining a very good values of correctness (93% - 97%) (see green dotted line in figure 1). This means that in noised condition the algorithm retrieves correctly the main part of the shape of the object but excludes some details, for example the hair (see top line of figure 4). The impact of this issue can be reduced making less resctrictive the threshold λ.

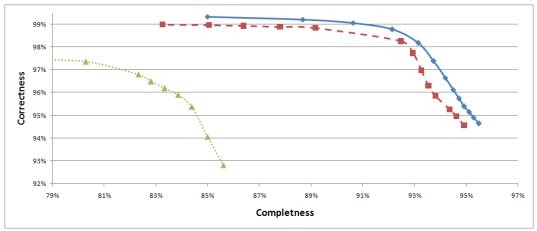


Figure 1: Correctness vs. completness for different sequences.

Firstly we introduce a new set of point A, defined by the following equation 6:

$$\{(I_y > \lambda) \vee [(I_y < \lambda) \wedge (I_{8n} > \delta * \lambda)], \delta \in [0,1]\} \rightarrow \{I_y \in A\} \quad (6)$$

where I_{8n} is the intensity of all the neighbours of the pixel y considering the 8-connection.

This set is obviously greater than the precedent, but has still some imprecision (see figure 2(b)). So we have designed another set of points, called M, applying to A a series of morphological operations (see figure 2(c)). The union of A e M produces a more precise set, L, that can fix the inaccuracies of A.

Figure 2(d) shows all the points of the distance map that have a correspondent intensity belonging to L. The comparison with figure 2(a), that presents the standard thresholding method, shows a notable increase of details with the new solution.

The appropriate sequence of morphological operations was established experimentally making tests with different light conditions and subjects. Generally the best results were obtained applying in order two dilation, five erosion and a final dilation.

$$L = A \cup M \quad (7)$$

We can now redefine as follow the equation 3, accordingly with the proposed corrections.

$$\{x \in C, S(x,y) < \theta, I_y \in L\} \rightarrow \{y \in C\} \quad (8)$$

The other parts of region growing do not need another modification and can be implemented like described previously.

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

Figure 2: A distance map filtered with different type of thresholding: (a) point with $I_y > \lambda$; (b) only set A with $\delta = 0.5$; (c) only set M; (d) set L.

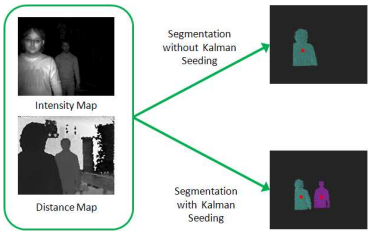


Figure 3: Segmentation without and with kalman seeding.

4 TRACKING

Many related works have examined the potential of Time-of-Flight cameras for tracking. The proposed approachs are very different: they range from an Expectation Maximization algorithm (Hansen et al., 2008) to a method based on depth distribution (Parvizi and Wu, 2008). The use of multiple cameras was investigated in (Guomundsson et al., 2008), but its applicability appears limited to a little ambient. Another interesting solution involves the integration of color and depth data for obtaining a more precise outcome (Bleiweiss and Werman, 2009).

We have experimented with a traditional Kalman filter to track the clusters. This kind of solution is already been analyzed in (Bevilacqua et al., 2006) with good results, but in that case the camera was placed as to provide a top-down view of the scene. This simplifies the elaboration but reduces significantly the visible area. Moreover all the detailes of the people are lost. So we always use a frontal view of the scene to made a more versatile implementation.

The Kalman state has six dimensions referring to centroid coordinates, i.e. (x, y, z, v_x, v_y, v_z), all expressed in image coordinates, as the SR3000 provides output data already organized in 3D Cartesian coordinates. After segmenting an image by region growing, we compare the detected clusters and those being tracked. The association between measured clusters and Kalman clusters is evaluated by minimum square euclidean distance between their centroids.

We compute a Gaussian representation of Kalman cluster at time $t-1$ and use its updated centroid position at time t to delineate the image region where the cluster should appear in frame t.

In case of cluster occlusion the kalman tries to estimate the more probable path of the disappeared cluster using its last detected movements and increasing the research area in order to compensate estimation error. If the cluster reappears shortly (at most within 30-40 frames) in a position closed to the predicted one it can be reassigned to its precedent kalman. On the contrary its kalman can be reinitialized and reassigned to a new cluster.

The data provided by our method can be used not only for tracking but also for increasing the accuracy of the segmentation, like presented in figure 3. In this situation we have two foreground objects to follow: if one of them gets too close to the sensor its intensity values grow too much and accordingly all the seeds will be concentrated on it, excluding the second object from region growing. This issue can be fixed using the informations generated by the prediction step of Kalman filter as new input for seeding phase. We seed at time t in all pixels around the predicted centroid at time $t+1$.

Adopting this procedure it is possible to extract and to track at the same time middleground and foreground objects.

5 EXPERIMENTAL RESULTS

5.1 Foreground Segmentation

In order to evaluate the robustness against noise of our segmentation algorithm we consider some video sequences acquired in our lab with artificial and natural illumination. The foreground objects extracted are visualized like 3D models where the vertexes positions are supplied by distance map and the color is obtained normalizing the intensity map. Figure 4 shows the results obtained with different subjects using the old (section 3.1) and the new method (section 3.2). After the improvements the extracted data contain a lot of more significant details, like hair or part of arms and legs (see bottom line of figure 4). This significant increase of completeness involves a small reduction of correctness, generates by new false positive at the border of the extracted objects. However these inaccuracies can be accepted considering the advantages supplied by a more complete model.

5.2 Performance Evalutation

We run our segmentation and tracking method on different models of computers in order to determine its

Figure 4: Segmentation results without (top) and with (bottom) the proposed noise correction.

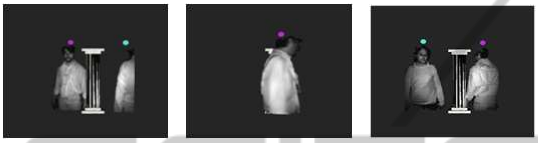


Figure 5: Tracking sequence.

Table 1: Speed on different computers: PC1: Intel Core 2 Quad Q9300 2.60 GHZ with Nvidia GeForce GTX 260; PC2: Intel Core 2 Duo T6600 2.40 GHz with Nvidia GeForce 9800 GTX; PC3: Intel Pentium IV 3.00 GHZ with Nvidia GeForce FX 5600; NB1: Intel Core 2 Duo L9400 1.86 GHZ with Integrated Graphic; NB2: Intel Pentium M 750 1.86 GHz with ATI Radeon X700.

Computers	Test 1	Test 2
Computer 1	18 fps	44 fps
Computer 2	18 fps	42 fps
Computer 3	18 fps	18 fps
Notebook 1	18 fps	28 fps
Notebook 2	18 fps	24 fps

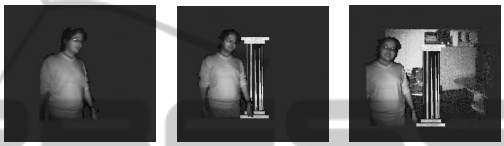
performances on low, middle and high level CPUs.

We considered two types of situations: online execution, with input directly provided by the camera, and offline execution, where the program analyzes precedent recorded video sequences. The first experiment is useful to test if it is possible using the SR3000 at its standard frame rate (18-20 fps as mentioned in section 2); while the second one gives us the maximum obtainable speed. This last data is particulary relevant considering the capabilities of the more recent versions of TOF cameras, like SR4000, that can reach 54 fps.

Table 1 summarizes the obtained results. The method turns out to be not computationally expensive, the first test is passed in all the cases and also the second shows very good performances. Also with a 10 years old processor (see computer 3 in table 1) we can reach the real-time execution. It is interesting to notice than the load for the GPU is very small for these operations, involves only the final visualization. So it possible to transfer the heaviest phases of the segmentation algorithm on graphic hardware for taking advantages of its parallel computational capabilities. This is a promising solution for increasing the performances, especially if we use a low level CPU.

5.3 Mixed Reality

The mixed reality is the merge of real and virtual worlds to produce a new environment where physical and digital objects coexist and interact in real-time. The Time-of-Flight camera are very useful in this kind of scenario (Bartczak et al., 2008), so it is a good choice for testing all the features of our system.



(a) (b) (c)

Figure 6: Mixed reality steps: (a) foreground segmentation; (b) adding of virtual object; (c) adding of 3D backgroud provided by TOF.

In an initial phase we considered only the data provided by SR3000, then we integrated the distance data with the color information supplied by a traditional webcam. This last test was made only for checking the feasibility of the color integration in our method, so we used a very low resolution webcam. In future improvements we will consider a high resolution camera for provide a more likeable outcome.

For solve the cameras calibration issue we studied the more recent approaches, as the ortophoto generation described in (Reulke, 2006) or the projective texture maps analyzed in (Lindner and Kolb, 2007). After some tests with both of them we chose a solution quite similar to the first one.

For both the cases (black and white and color) we follow the same procedure summarized in figure 6: firstly we extract the foreground subject, then add the column and finally insert the backround, that is simply the 3D distance map generated by the TOF. Some examples results with color are shown in figure 7.

Figure 7: Mixed Reality with color.

The system is able to manage the interaction of multiple clusters and also the tracking appears to be

robust to the occlusions. This feature is better explained by figure 5 where there are displayed different frames of the same sequence. The colored spheres on the top of the two subjects are the marks of kalman trackers, note how the correspondence between cluster and its kalmans is always mantained. The tracking data can be useful also for increasing the realism of the simulation, for example with a real time adjustment of the position of the virtual objects. A similar solution was firstly described in (Bartczak et al., 2008) but suffers of some limitations: it need an offline computation of the background model and it can discriminate only one subject at time.

In accordance with the description of section 3 our approach appears to be more general. We do not need any preprocessing operation except for the calibration and all the steps of segmentation are executed in real-time without any information on the background and with heterogeneus types of illumination.

6 CONCLUSIONS

We have presented a new approach to multiple subjects segmentation and tracking, that exploits the intrinsic characteristics of the intensity and distance signals generated by modulated-light TOF. Our method is able to reduce the effect of noise introduced by sun light interferences, through a flexible intensity thresholding and the mathematical morphology.

The experimental results show that the proposed approach can be used in multimedia applications, like mixed reality.

The performance tests prove that the system is computationally efficient and can reach real-time execution also with low level computers. A future parallel implementation of the most power intensity parts of the system, like mathematical morphology, can further increase the performances.

Other improvements include data fusion of color and TOF cameras for a more robust segmentation.

REFERENCES

Bartczak, B., Schiller, I., Beder, C., and Koch, R. (2008). Integration of a time-of-flight camera into a mixed reality system for handling dynamic scenes, moving viewpoints and occlusions in real-time. In *3DPVT08, Fourth International Symposium on 3D Data Processing, Visualization and Transmission*.

Bevilacqua, A., Stefano, L. D., and Azzari, P. (2006). People tracking using a time-of-flight depth sensor. In *AVSS 06, Video and Signal Based Surveillance*. IEEE Computer Society.

Bianchi, L., Dondi, P., Gatti, R., L.Lombardi, and Lombardi, P. (2009). Evaluation of a foreground segmentation algorithm for 3d camera sensor. In *ICIAP 2009, 15th International Conference of Image Analysis and Processing*. Springer.

Bleiweiss, A. and Werman, M. (2009). Real-time foreground segmentation via range and color imaging. In *Dyn3D09, Proceedings of DAGM2009 Workshop on Dynamic 3D Imaging*. Springer.

Crabb, R., Tracey, C., Puranik, A., and Davis, J. (2008). Real-time foreground segmentation via range and color imaging. In *CVPRW 2008, Computer Vision and Pattern Recognition Workshops*. IEEE Computer Society.

Guomundsson, S., Larsen, R., Aanaes, H., Pardas, M., and Casas, J. R. (2008). Tof imaging in smart room environments towards improved people tracking. In *CVPRW 2008, Computer Vision and Pattern Recognition Workshops*. IEEE Computer Society.

Hansen, D. W., Hansen, M. S., Kirschmeyer, M., Larsen, R., Silvestre, D., and Silvestre, D. (2008). Cluster tracking with time-of-flight cameras. In *CVPRW 2008, Computer Vision and Pattern Recognition Workshops*. IEEE Computer Society.

Kolb, A., Barth, E., Koch, R., and Larsen, R. (2010). Time-of-flight cameras in computer graphics. In *Computer Graphics Forum volume 29, issue 1*. Wiley.

Lindner, M. and Kolb, A. (2007). Data-fusion of pmd-based distance-information and high-resolution rgb-images. In *ISSCS 2007, International Symposium on Signals, Circuits and Systems*.

Oggier, T., Lehmann, M., Kaufmann, R., Schweizer, M., Richter, M., Metzler, P., Lang, G., Lustenberger, F., and Blanc, N. (2004). An all-solid-state optical range camera for 3d real-time imaging with sub-centimeter depth resolution (swissranger). In *SPIE 2004, Society of Photo-Optical Instrumentation Engineers Conference Series*.

Oprisescu, S., Falie, D., Ciuc, M., and Buzuloiu, V. (2007). Measurements with tof cameras and their necessary corrections. In *ISSCS 2007, International Symposium on Signals, Circuits and Systems*.

Parvizi, E. and Wu, Q. J. (2008). Multiple object tracking based on adaptive depth segmentation. In *Canadian Conference of Computer and Robot Vision*, pages 273–277. IEEE Computer Society.

Reulke, R. (2006). Combination of distance data with high resolution images. In *IEVM06, Image Engeeniring and Vision Metrology*.

Santrac, N., Friedland, G., and Rojas, R. (2006). High resolution segmentation with a time-of-flight 3d-camera using the example of a lecture scene. Technical report, http://www.inf.fu-berlin.de/inst/ag-ki/eng/index.html.