Robust Volumetric Reconstruction from Noisy Multi-view Foreground Occupancy Masks

Fan Chen* and Christophe De Vleeschouwer†
* Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan
E-mail: chen-fan@jaist.ac.jp
† Universit catholique de Louvain, Louvain-la-neuve, Belgium
E-mail: Christophe.Devleeschouwer@uclouvain.be

***Abstract*—We recover the shape of a moving object, represented by the occupancy of grid voxels in the physical space, from the consistency of its foreground occupancy in multiple camera views. To deal with the noise due to loose temporal synchronization, lighting variation and calibration errors, we integrate the foreground occupancy of each grid voxel over its local neighborhood. Especially, we design a way to decouple each integral direction, so that the accumulation along this direction can be efficiently computed by integrating pixels along a row/column in the transformed foreground masks, which also enables the access to the intermediate physical positions between two neighboring voxels. We further accelerate this integration by using the technique of integral images. The performance of the proposed method has been investigated experimentally.**

I. Introduction

Shape reconstruction is not only meaningful in shape measurement, but also useful in providing clues (e.g. skeleton) to activity detection in various entertainment/surveillance applications. Many methods have been proposed for shape reconstruction from multiple photographies[1], including volumetric reconstruction using voxels[2], polyhedral visual hull using tangent surfaces [3], and space carving via photo consistency. Volumetric reconstruction is popular due to its simple and flexible presentation of 3D shapes, but is usually computationally heavy. Although less computationally expensive, visual hull-based methods require more cameras to make a surface accurate and are more sensitive to noisy foregrounds, which limits their potential applications. Furthermore, we avoid adding extra restrictions to camera placement, lighting conditions, or grayscale/color cameras. Hence, methods based on color consistency, e.g. [2], are not considered here.

A simple but efficient way of volumetric reconstruction is to accumulate the foreground occupancies, following the homographic occupancy constraint in [4], which explores that only occupied voxels can be correctly warpped into the foreground of all camera views. However, this consistency is usually biased in real applications, due to loose temporal sychronization and inaccurate calibration. Furthermore, the foreground occupancy mask usually contains noise due to the inefficiency of automatic background removal methods. In Fig.1, we present some sample noises in the foreground occupancy mask obtained using the state-of-the-art background extractor, under different tolerance thresholds, including

- Noise due to varying light conditions (e.g., the horizontal noise band due to the strobe lamps here) or camera CCD noises (, which is not obvious here, but is severe, e.g., on a surveillance video taken in early morning or evening, or in night view mode). This kind of noise could be even more significant if the background template were not dynamically updated. However, this kind of noise is less consistent in different camera views;
- Noise due to foreground/background texture patterns. This pattern could either be the original pattern, e.g., the zebra patterning of the sweater, or a pattern generated by the movement of body parts, e.g., the wrinkles of the cloths. Except for the case that the object intends to hide himself by taking a very similar pattern to the background, this kind of noise will not cause a large area of data missing in the foreground occupancy mask;
- Noise due to shadows. This is not really a noise, because the shadow is something that does exist in the environment, with well consistency in multiple camera views;
- Noise due to an over-high tolerance threshold or a similar pattern between the foreground and background. This kind of noise usually results in large area data missing of foreground occupancy masks, e.g., the legs here.

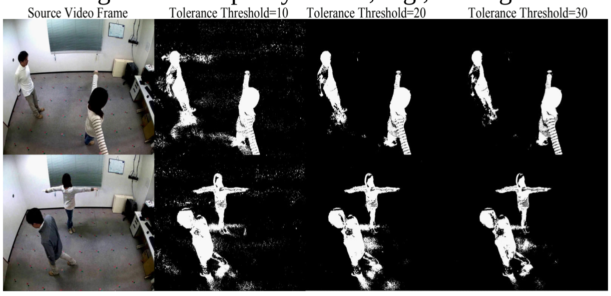

Fig. 1. Some sample noises in the foreground occupancy mask obtained using the state-of-the-art background extractor, under different tolerance thresholds.

For object tracking, it could be better to raise the tolerant threshold for a neat foreground mask, so as to suppress the noises from shadows and varying lighting conditions [6]. For recovering the body shape, it is safer to preserve more potential foreground pixels, than taking the risk of losing important body parts. Furthermore, background noises under a low threshold are usually inconsistent in different camera views, which thus could be filtered out via validating the homographic occupancy constraint. Hence, we prefer more to reconstruct a shape from a noisy foreground mask.

Robust reconstruction against this noise is further improved by considering its neighboring points in the physical 3D space. Due to projective transformation and lens distortion, it is inefficient to apply the mean filter directly on foreground masks. [5] corrects the vanishing point in Z-direction to ensure that people are always standing vertically in the image plane, so as to easily apply a rectangle bounding-box on the images for people tracking. In the present paper, we extend this correction to all three directions and design an efficient spatial integration around each grid voxel, based on which we propose a more robust volumetric shape reconstruction method.

In the following parts, we first explain how we could achieve efficient spatial integration around each grid voxel in Section II, and then introduce the overall method for shape reconstruction in Section III. We then provide experimental results in Section IV and conclude the paper in Section V.

II. Efficient Spatial Integration

Spatial integration over a voxel's neighbourhood improves both the robustness against noisy foregrounds, and the efficiency of utilizing the foreground information. We achieve the spatial integration by decoupling each integral direction, which is a process to transform the integral direction into an efficient scanning order in the foreground occupancy masks. As shown in Fig.2, this process includes two major steps: 1) We convert spatial integration into row/column-wise accumulation of foreground occupancy; 2) We perform row/columnwise equalization of integral length, so as to determine the integral range in the foreground masks. Here, we only explain the processing of the X-direction integration. Derivation of transformations in other directions is similar.

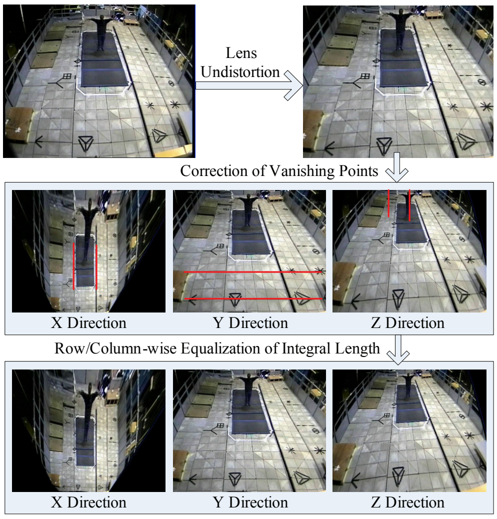

Fig. 2. Diagram for decoupling the X, Y, and Z integral directions in the original camera view via vanishing point correction and image shearing.

A. Correction of vanishing points

Given a camera where the lens distortion has been corrected, projection matrix $\mathbf{P}$ projects a physical point $\mathbf{p} = [XYZ1]^T$ into an image point $\mathbf{i} = [uv1]^T$, i.e., $\mathbf{i} = \mathbf{Pp}$. We correct the vanishing point, so that all paralleled lines along the X-direction are projected into paralleled vertical lines in the camera view. This is achieved by multiplying $\mathbf{P}$ with a 3×3 matrix $\mathbf{B}$, i.e., the corrected projection matrix $\hat{\mathbf{P}} = \mathbf{BP}$, and

$$\mathbf{B} = \begin{bmatrix} 1 & \text{-}\mathbf{P}_{11}/\mathbf{P}_{21} & 0 \\ 0 & 1 & 0 \\ 0 & \text{-}\mathbf{P}_{31}/\mathbf{P}_{21} & 1 \end{bmatrix}. \tag{1}$$

Q_{ij} denotes the element in the i^{th} row j^{th} column of Matrix Q. It is easy to verify that both $\hat{\mathbf{P}}_{11}$ and $\hat{\mathbf{P}}_{31}$ are zero, and thus any changing in the X-th direction will only be reflected by the vertical direction in the modified camera view.

B. Row/column-wise equalization of integral length

We further left multiply $\hat{\mathbf{P}}$ with matrix $\mathbf{S}$, which reads

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 \\ \text{s} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{2}$$

with $s = (\hat{\mathbf{P}}_{23}\hat{\mathbf{P}}_{32} - \hat{\mathbf{P}}_{22}\hat{\mathbf{P}}_{33})/(\hat{\mathbf{P}}_{12}\hat{\mathbf{P}}_{33} - \hat{\mathbf{P}}_{13}\hat{\mathbf{P}}_{32})$. Projection matrix $\mathbf{S}\hat{\mathbf{P}}$ projects $\mathbf{p} = [XYZ1]^T$ into $\hat{\mathbf{i}} = [\hat{u}\hat{v}1]^T$, where

$$\hat{u} = \frac{\hat{\mathbf{P}}_{12}Y + \hat{\mathbf{P}}_{13}Z + \hat{\mathbf{P}}_{14}}{\hat{\mathbf{P}}_{32}Y + \hat{\mathbf{P}}_{33}Z + \hat{\mathbf{P}}_{34}}, \tag{3}$$

$$\hat{v} = \frac{\hat{\mathbf{P}}_{21}X + Const_1}{\hat{\mathbf{P}}_{32}Y + \hat{\mathbf{P}}_{33}Z + \hat{\mathbf{P}}_{34}} + Const_2. \tag{4}$$

In the integral image whose X direction has been corrected using the process in Fig.2, the integral length $l^{\mathrm{X}}(\mathbf{p}_1, \mathbf{p}_2)$ between two points $\mathbf{p}_1 = [X_1YZ1], \mathbf{p}_2 = [X_2YZ1]$ is linearly proportional to $X_1 - X_2$, i.e.,

$$l^{\mathrm{X}}(\mathbf{p}_1, \mathbf{p}_2) = \frac{\hat{\mathbf{P}}_{21}(X_1 - X_2)}{\hat{\mathbf{P}}_{32}Y + \hat{\mathbf{P}}_{33}Z + \hat{\mathbf{P}}_{34}}, \tag{5}$$

which enables us to easily compute the integral range.

From Eq.(4), any points in the YZ homography plane that are projected into the same row of the image, have the same unit integral length. (Here, the unit integral length at a voxel is defined as the pixel distance corresponding to a unit length from this voxel in the physical space.) This property can be used to improve the computational efficiency.

III. Robust Shape Reconstruction

As shown in Fig.3, our reconstruction method with spatial integration consists of two major steps:

Step 1) We prepare integral images of foreground occupancy masks, by extracting foreground masks, correcting vanishing points and computing integral images;

Step 2) We determine the integral range in each integral image for the current voxel. The integral origin can be computed from its distance to a referential plane, while the unit integral length only depends on its projected position in this referential plane (Eq.5).

We then calculate the confidence $\mathcal{C}([XYZ1]^T)$ of voxel occupancy by accumulating the foreground occupancy masks

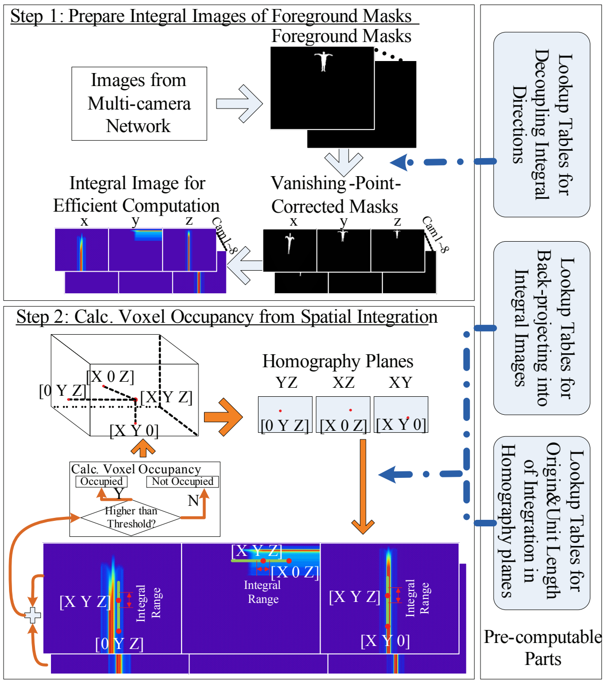
Step 1: Prepare Integral Images of Foreground Masks
Foreground Masks
Images from Multi-camera Network
Integral Image for Efficient Computation
Vanishing-Point-Corrected Masks
x
y
z
Cam1-8
Step 2: Calc. Voxel Occupancy from Spatial Integration
[X 0 Z]
[0 Y Z]
[X Y Z]
[X Y 0]
Homography Planes
YZ
XZ
XY
[0 Y Z]
[X 0 Z]
[X Y 0]
Calc. Voxel Occupancy
Occupied
Not Occupied
Y
N
Higher than Threshold?
[X Y Z]
Integral Range
[0 Y Z]
[X Y Z]
[X 0 Z]
Integral Range
[X Y Z]
Integral Range
[X Y 0]
Lookup Tables for Decoupling Integral Directions
Lookup Tables for Back-projecting into Integral Images
Lookup Tables for Origin&Unit Length of Integration in Homography planes
Pre-computable Parts

Fig. 3. The block diagram of the proposed method of shape reconstruction.

in the three space directions from all camera views, i.e.,

$$
\begin{aligned}
\mathcal{C}([XYZ1]^T) = \sum_c \{ & \int_{X-\Delta_X}^{X+\Delta_X} \mathcal{F}_c([xYZ1]^T)dx \\
& + \int_{Y-\Delta_Y}^{Y+\Delta_Y} \mathcal{F}_c([XyZ1]^T)dy \\
& + \int_{Z-\Delta_Z}^{Z+\Delta_Z} \mathcal{F}_c([XYz1]^T)dz\}, \qquad (6)
\end{aligned}
$$

where $\mathcal{F}_c([XYZ1]^T)$ is the corresponding value of point $[XYZ1]^T$ in the foreground occupancy mask. $\Delta_X, \Delta_Y, \Delta_Z$ are prespecified neighbourhood size of spatial integration in the X, Y, Z direction respectively. Note that we perform integration along three lines rather than over the whole cubic neighbourhood, so as to reach a balance between computational burden and robustness. $\Delta_X, \Delta_Y, \Delta_Z$ should be small with respect to the thickness of the object, so as to avoid losing details. We then compare this confidence value to a prespecified threshold to determine the voxel occupancy.

In order to improve the computational efficiency, we organize all pre-computable parts together, including the Look-Up Tables (LUT) for projecting the lens-distorted foreground masks to the integral image, LUTs for projecting three homography planes to their corresponding integral images, and LUTS for saving the origin position and unit integral length for each point in those referential homography planes.

IV. Experimental Results

We investigate the performance by using both the Muhavi database[7] and our own video database. Some sample images are shown in Fig.4. We first present some quantitative experimental results on artificial noises in Section IV-A, and then give some results on real noisy foreground masks.

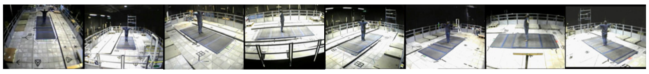

Fig. 4. Samples images of the MuHavi Database.

A. Results on robustness against artificial noises

Since the real noise is difficult to evaluate, we apply artificial noises to manually annotated silhouettes to simulate the noisy foreground masks, where the noise level is defined as the probability of flipping a foreground occupancy. We compare our method, which in fact applies adaptive mean-filtering on the image plane, to both reconstruction without smoothing and reconstruction from pre-smoothed foreground masks after applying fixed-size mean-filtering.

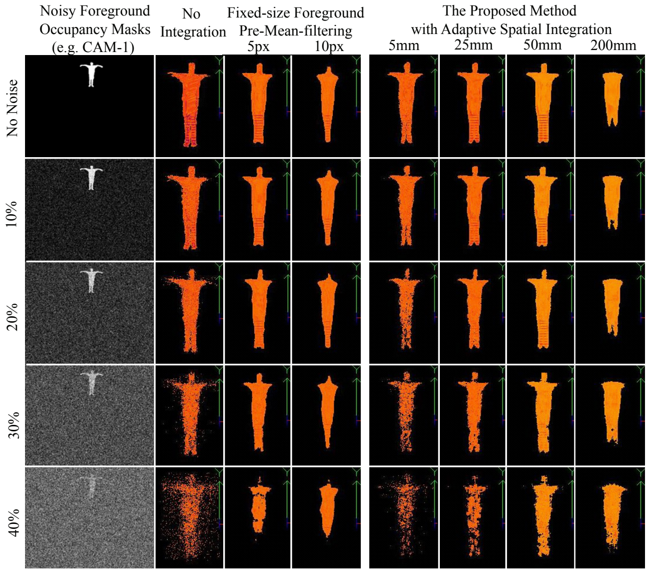
Noisy Foreground Occupancy Masks (e.g. CAM-1)
No Integration
Fixed-size Foreground Pre-Mean-filtering
5px
10px
The Proposed Method with Adaptive Spatial Integration
5mm
25mm
50mm
200mm
No Noise
10%
20%
30%
40%

Fig. 5. Reconstruction results are shown for seven cases, under different noise levels. The proposed method not only has less false positive alarms, but also has lower false negative alarms as shown in Fig.6.

In Fig.5, we compare seven cases, namely no integration, two cases on pre-smoothed foreground masks with their half window size being $5px$ and $10px$, four with spatial integration at length $\Delta_X = \Delta_Y = \Delta_Z = 5mm$, $\Delta_X = \Delta_Y = \Delta_Z = 25mm$, $\Delta_X = \Delta_Y = \Delta_Z = 50mm$, and a very large length $\Delta_X = \Delta_Y = \Delta_Z = 200mm$,, under five different noise levels from no noise to 40%. In the left column, we show the sample noisy foreground, while in the next columns, we display the resultant images of the reconstructed shapes. We have the following major observations:

1) With the increasing of integration length, the shape is better preserved under a higher noise level.

2) Spatial integration is useful in reducing the false positive alarms caused by isolated voxels, as a result of performing the adaptive mean filtering.

3) Under a high noise level, legs are missing in the results from fixed-size mean-filtering. In all camera views of the current database, legs are further than the upper body, which thus are thinner in the foreground masks. Under the fixed size mean-filtering, legs receive a stronger smoothing than the upper body, which made them closer to the background area and caused them disappearing after applying the threshold. This is not simply solvable by changing globally the window size of the filter. To the contrary, the legs are well preserved

in the proposed methods, which is considered to be an advantage of using adaptive window size in the pixel domain (to perform constant length integration in the physical world). This improvement is essential in some applications, e.g., when the recovered voxels are used later to skeletonize the body.

4) A proper integral length improves the robustness against strong noises, while an overlarge integral length loses details (e.g., the case of $200mm$), where a balance need to be found.

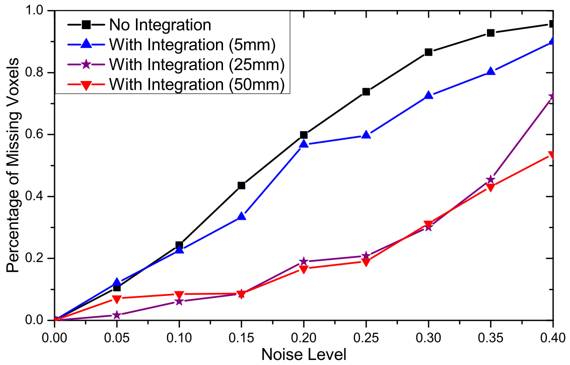

Fig. 6. Percentage of missing voxels over all positive grid voxels in the noise-free case has been plotted for four cases, which shows that reconstruction with space integration has less missing voxels under the same noise level.

Lacking of the ground-truth, it is difficult to check the amount of missing voxels due to the noise. For each case with a different integral length, we approximate its ground-truth by the shape obtained from the noise-free foreground masks, and plot the percentage of missing voxels WRT the ground-truth in Fig.6. It shows that reconstruction with spatial integration has also a lower number of missing voxels under the same noise level. For the case of $5mm$, it is close to the case without integration, because $5mm$ is relatively small in pixel sizes due to the low resolution (720px × 576px) of the camera view, while integral length $50mm$ is already quite large and thus achieves little improvement over that of $25mm$.

B. Results on robustness against real noises

We are validating the performance against real noises, on our own video dataset. In Fig.7, we compare two cases, i.e., reconstruction with and without spatial integration, under different thresholds of voxel filtering, and present the results on five typical frames. Here, the integral length is set to $50mm$. We make the following observations:

- In order to preserve the body parts, we adopt a sensitive foreground extractor, which results in noisy foreground masks. Since the noise is in general less consistent than body parts in different camera views, most of noisy voxels are removable by setting a proper threshold.
- A very high threshold as 0.9 is risky in losing key body parts. Spatial integration enlarges the difference between isolated noises and those noisy body parts, which further suppresses noises and separates objects better under low filtering thresholds. As a result, spatial integration provides a wider range of working thresholds, which could improve the overall quality and stability of reconstruction over a large set of frames.

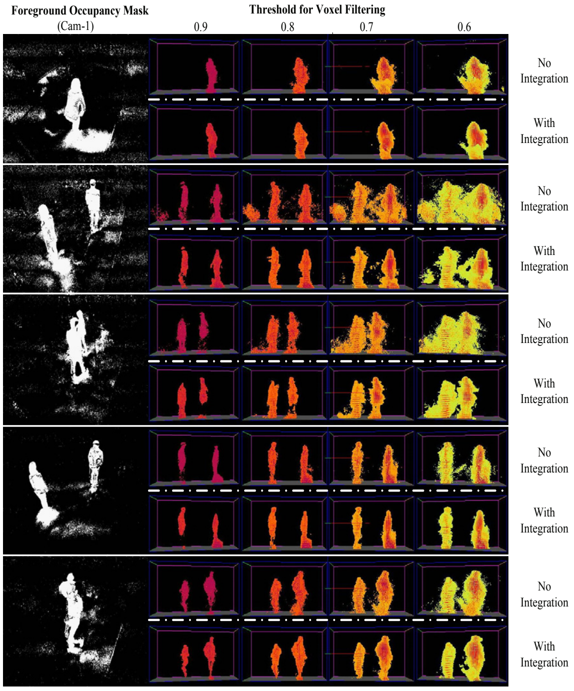

Fig. 7. Some results on real noises. Spatial integration helps to reduce the noisy voxels in the final results, and thus makes the result more robust under different threshold setup of voxel filtering.

V. Conclusion

We propose a robust method for recovering volumetric shapes from multi-view videos, where the occupancy of grid voxels in the 3D space is determined by accumulating its neighboring points in all foreground occupancy masks. There are two major contributions: firstly, we designed an efficient way to perform this spatial integration; secondly, we develop a reconstruction method using this integration technique. From experimental results, we confirm that the proposed method is efficient in both filling missing voxels and removing isolated false voxels, which thus has a higher robustness against noisy foreground occupancy masks. The possibility of applying adaptive processing regarding the boundary of objects for maintaining local details during smoothing will be discussed in our future work. Although the trend was evaluated in this paper, we lacks of ground truth of shape data for precision evaluation, which will be investigated in our future work, along with body skeletonization and action recognition.

Acknowledgment

The JAIST's contribution has been partly funded by the Japanese Grants-in-Aid for Scientific Research (Start-up Support) No.22800026 and the Japanese MEXT Grant-in-Aid for Young Scientists (B) No.23700110.

References

[1] Aghajan H., and Cavallaro A., "Multi-camera Networks,"Elsevier,2009.

[2] Dyer C.R., "Volumetric Scene Reconstruction From Multiple Views," Foundations of Image Understanding, Kluwer, 469-489, 2001.

[3] Kehl R., Bray M., and Luc Van Gool, "Full Body Tracking from Multiple Views Using Stochastic Sampling," CVPR, 2:129-136, 2005.

[4] Khan S.M.,and Shah M., "Tracking multiple occluding people by localizing on multiple scene planes," IEEE Trans. PAMI, 31:505-519, 2009.

[5] Yao J., and Odobez J.-M., "Multi-camera multi-person 3D space tracking with MCMC in surveillance scenarios, " ECCV 2008 (M2SFA2), 2008.

[6] Delannay D., Danhier N., and De Vleeschouwer C., "Detection and recognition of sports (wo)men from multiple views," *ICDSC'09*, 2009.

[7] MuHAVi Database, http://dipersec.king.ac.uk/MuHAVi-MAS/