Probabilistic visibility for multi-view stereo

Carlos Hernández[1]
Computer Vision Group
Toshiba Research Europe

George Vogiatzis[2]
Computer Vision Group
Toshiba Research Europe

Roberto Cipolla[3]
Dept. of Engineering
University of Cambridge

{carlos.hernandez[1]|george.vogiatzis[2]}@crl.toshiba.co.uk, cipolla@eng.cam.ac.uk[3]

Abstract

We present a new formulation to multi-view stereo that treats the problem as probabilistic 3D segmentation. Previous work has used the stereo photo-consistency criterion as a detector of the boundary between the 3D scene and the surrounding empty space. Here we show how the same criterion can also provide a foreground/background model that can predict if a 3D location is inside or outside the scene. This model replaces the commonly used naive foreground model based on ballooning which is known to perform poorly in concavities. We demonstrate how the probabilistic visibility is linked to previous work on depth-map fusion and we present a multi-resolution graph-cut implementation using the new ballooning term that is very efficient both in terms of computation time and memory requirements.

1. Introduction

Digital modeling of 3D objects is becoming increasingly popular and necessary for a wide range of applications such as cultural heritage preservation, online shopping or computer games. Although laser range scanning remains one of the most popular techniques of acquiring shape, the high cost of the equipment, complexity, and difficulties to capture color are three big disadvantages. As opposed to laser techniques, image-based techniques provide an efficient and easy way to acquire shape and color by simply capturing a sequence of photographs of the object. Among the vast literature available on image-based modeling techniques, recent work based on volumetric approaches shows an excellent compromise between computation time and accuracy. However, these methods exhibit very strong memory requirements and are difficult to tune in order to obtain the best results.

In this paper we present a new formulation of the multi-view stereo problem that closely resembles a 3D segmentation. The new formulation is both computationally and memory efficient. The main contribution of this paper is that it is explained how the photo-consistency criterion can define the *outside* of an object. It upgrades previous reasoning based on photo-consistency where only the notion "*on the object*" is available, but inside and outside are indistinguishable. The existence of a background/foreground distinction is crucial for volumetric methods since, whenever it is not available, an ad-hoc model is introduced, usually in the form of an inflationary or deflationary force, *i.e.*, a ballooning term [6]. The negative effect of this new term is that it also acts as a regularization term and can make details or thin structures disappear from the final reconstruction. Compared to previous work, our formulation provides a foreground/background model that is *data-aware*. This model can be used as an intelligent ballooning term providing better reconstruction results. The new background model can be used with previous existing techniques for 3D segmentation such as level-sets or graph-cuts while improving their accuracy and simplicity for the task of 3D reconstruction.

This paper is organized as follows: In Section 2 we review the literature and explain the main motivation of the paper. In Section 3 we sketch the basics of the technique and describe how to use the concept of probabilistic visibility as an intelligent ballooning term in a multi-resolution graph-cut implementation. Section 4 describes the theory behind the new probabilistic approach to visibility computation from a set of depth-maps. Finally in Section 5 we validate the proposed approach by showing high quality reconstructions and comparing it to previous algorithms.

2. Motivation and related work

We consider a set of input photographs of an object and their corresponding camera projection matrices. The object is considered to be sufficiently textured so that dense correspondence can be obtained between two different images. The goal is to obtain a 3D surface representation of the object, *e.g.*, a triangular mesh.

As recently reviewed by Seitz et al. [18], a vast quantity

of literature exists on how to obtain such a 3D representation from images only. Nearly all of them use a photo-consistency measure to evaluate how consistent is a reconstruction with a set of images, *e.g.*, normalized cross correlation or sum of square differences.

The inspiration of this paper is recent progress in multi-view stereo reconstruction, and more specifically on volumetric techniques. The overview of state-of the-art algorithms for multi-view stereo reconstruction reported by Seitz et al. [18] shows a massive domination of volumetric methods. Under this paradigm, a 3D cost volume is computed, and then a 3D surface is extracted using tools previously developed for the 3D segmentation problem such as snakes [12], level-sets [17] or more recently graph-cuts [21, 19, 4, 9, 13, 16, 20].

The way volumetric methods usually exploit photo-consistency is by building a 3D map of photo-consistency where each 3D location gives an estimate of how photo-consistent would be the reconstructed surface at that location. The only requirement to compute this photo-consistency 3D map is that camera visibility is available. Some methods use an initial approximation of the true surface to estimate visibility, such as the visual hull [21]. Iterative methods use instead the notion of "current surface". The visibility computed from the reconstructed surface at iteration $i-1$ is then used to compute photo-consistency at iteration i, improving the reconstruction gradually [8]. Finally, some recent methods are able to compute a "*visibility-independent*" photo-consistency where occlusion is treated as an additional source of image noise [12].

Independently of how visibility is computed, all the volumetric methods suffer from a limitation that is specific to the multi-view stereo problem, namely that there is no straightforward way of defining a foreground/background model for the 3D segmentation problem. This is because in this problem our primary source of geometric information is the *correspondence cue* which is based on the following observation: A 3D point located *on* the object surface projects to image regions of *similar* appearance in all images where it is not occluded. Using this cue one can label 3D points as being *on* or *off* the object surface but cannot directly distinguish between points *inside* or *outside* it. This lack of distinction has been historically solved by using a data-independent *ballooning* term that produces a constant inflationary tendency. The motivation for this type of term in the active contour domain is given in [6], but intuitively, it can be thought of as a shape prior that favors objects that fill the bounding volume in the absence of any other information. On one hand, if the ballooning term is too large, then the solution tends to over-inflate, filling the entire bounding volume. On the other hand, if it is too small, then the solution collapses into an empty surface.

In this paper we propose a probabilistic framework to construct a data-aware ballooning term from photo-consistency only. This framework is related to the work of [5, 1, 10] in that we all aim to model geometric occlusion in a probabilistic way. However we are the first to study the problem in a volumetric framework adapted to the 3D segmentation problem. Roughly speaking, instead of just assigning a photo-consistency value to a 3D location, this value is also propagated towards the camera centers that where used to compute it. The key observation about photo-consistency measures is that, besides providing a photo-consistency score of a 3D particular location, they also give additional information about the space *between* the location and the cameras used to compute the consistency. In other words, if a set of cameras gives a high photo-consistency score to a 3D location, they give at the same time a "background" score of the same strength to the 3D segments linking the 3D location with the camera centers. This follows from the fact that, if the surface was really at that location, then the cameras that gave it a high photo-consistency score would indeed see the surface without occlusion, *i.e.*, the segments linking the camera centers with the 3D location would all be background.

To our knowledge, this "background" score is not used in any volumetric multi-view stereo algorithm, perhaps with the exception of [11], where photo-consistency is used to generate depth-maps which are then merged together using a volumetric depth-map fusion technique [7]. It turns out that depth-map fusion techniques are very related to our probabilistic approach. In fact, we demonstrate how our probabilistic approach explains and generalizes the work of Levoy and Curless [7] on using signed distance functions for depth-map fusion. As we show in Section 4, the work of [7] naturally arises as the probabilistic solution whenever the sensor noise is modeled with a logistic distribution.

3. Multi-view stereo using multi-resolution graph-cuts

In [3] and subsequently in [2] it was shown how graph-cuts can optimally partition 2D or 3D space into 'foreground' and 'background' regions under any cost functional consisting of the following two terms:

- **Labeling cost:** for every point in space there is a cost for it being labeled 'foreground' or 'background'.
- **Discontinuity cost:** for every point in space, there is a cost for it lying on the boundary between the two partitions.

Mathematically, the cost functional described above can be seen as the sum of a weighted *surface area* of the boundary surface and a weighted *volume* of the 'foreground' region

as follows:

$$E[S] = \int_S \rho(\mathbf{x}) dA + \int_{V(S)} \sigma(\mathbf{x}) dV \qquad (1)$$

where S is the boundary between 'foreground' and 'background', $V(S)$ denotes the 'foreground' volume enclosed by S and ρ and σ are two scalar density fields. The application described in [3] was the problem of 2D/3D segmentation. In that domain $\rho(\mathbf{x})$ is defined as a function of the image intensity gradient and $\sigma(\mathbf{x})$ as a function of the image intensity itself or local image statistics.

This model balances two competing terms: The first one minimizes a surface integral of photo-consistency while the second one maximizes the volume of regions with a high evidence of being foreground. While the photo-consistency term is relatively easy to compute from a set of images, very little work has been done to obtain an appropriate ballooning term. In most of the previous work on volumetric multi-view stereo the ballooning term is a very simplistic inflationary force that is constant in the entire volume, i.e., $\sigma(\mathbf{x}) = -\lambda$. This simple model tries to recover thin structures by maximizing the volume inside the final surface. However, as a side effect, it also fills in concavities behaving as a regularization force and smoothing fine details.

When silhouettes of the object are available, an additional *silhouette cue* can be used [21], which provides the constraint that all points *inside* the object volume must project inside the silhouettes of the object. Hence the silhouette cue can provide some foreground/background information by giving a very high likelihood of being *outside* the object to 3D points that project outside the silhouettes. However this ballooning term is not enough if thin structures or big concavities are present, in which case the method fails (see Fig. 4 middle row). Very recently, a data driven, foreground/background model based on the concept of *photo-flux* has been introduced [4]. However, the approach requires approximate knowledge of the object surface orientation which in many cases is not readily available.

Ideally, the ballooning term should be linked to the notion of visibility, where points that are not visible from any camera are considered to be inside the object or **foreground**, and points that are visible from at least one camera are considered to be outside the object or **background**. An intuition of how to obtain such a ballooning term is found in a classic paper on depth sensor fusion by Curless and Levoy [7]. In that paper the authors fuse a set of depth sensors using signed distance functions. This fusion relies on the basic principle that the space between the sensor and the depth map should be empty or background, and the space after the depth map should be considered as foreground. Here we propose to generalize this visibility principle and compute a probabilistic version of it by calculating the "*evidence of visibility*" from a given set of depth-maps and use it as an intelligent ballooning term.

The outline of the full system is as follows:

- create a set of depth-maps from the set of input calibrated images,
- derive discontinuity cost $\rho(\mathbf{x})$ from the set of depth-maps, *i.e.*, compute the photo-consistency term,
- derive labeling cost $\sigma(\mathbf{x})$ from the set of depth-maps, *i.e.*, use a data-aware ballooning term computed from the evidence of visibility and,
- extract the final surface as the global solution of the min-cut problem given $\rho(\mathbf{x})$ and $\sigma(\mathbf{x})$.

It is worth noting that the algorithm just described can also be used when the input is no longer a set of images but a set of depth-maps obtained from other types of sensor, *e.g.*, laser scanner. In this case, the system just skips the first step, since the depth-maps are already available, and computes ρ and σ directly from the set of depth-maps given as input.

3.1. Depth-map computation from images

The goal of this section is to generate a set of depth-maps $D_1, ..., D_N$ from a a sequence of images $I_1, ..., I_N$ calibrated for camera pose and intrinsic parameters. Each depth map is similar to a 2D image but where each pixel measures depth of the scene away from the sensor instead of a color value. In order to create the depth-maps from the set of input images, we propose to use a robust photo-consistency metric similar to the one described in [12] that does not need any visibility computation. This choice is motivated by the excellent results obtained by this type of photo-consistency metric in the recent comparison of 3D modeling techniques carried out by [18]. Basically, occlusion is considered as another type of image noise and is handled robustly in the same way as the lack of texture or the presence of highlights in the image. For a given image I_i, the depth $D_i(\mathbf{x})$ along the optic ray generated by a 3D location $\mathbf{x}$ is computed as follows:

- Compute the corresponding optic ray

$$\mathbf{o}_i(d) = \mathbf{x} + (\mathbf{c}_i - \mathbf{x})d \qquad (2)$$

  that goes through the camera's optic center $\mathbf{c}_i$ and the 3D location $\mathbf{x}$,

- As a function of the depth along the optic ray d, project the 3D point $\mathbf{o}_i(d)$ into the M closest cameras and compute M correlation scores between each neighbor image and the reference image using normalized cross correlation,

- combine the M correlation scores into a single score $\mathcal{C}(d)$ using a voting scheme as in [12], and find the final depth D_i as the global maximum of $\mathcal{C}$. The confidence on the depth D_i is simply $\mathcal{C}(D_i)$. As an optional test, a minimum confidence value can be used to reject depth estimations with very low confidence. The 3D locations of depths that are obtained along with their corresponding confidence are stored.

3.2. Discontinuity cost from a set of depth-maps

Once we have computed a depth-map for every input image, we can build the discontinuity map $\rho(\mathbf{x})$ for every 3D location $\mathbf{x}$. We propose a very simple accumulation scheme where for every 3D point $\mathbf{x}$ its total photo-consistency $\mathcal{C}(\mathbf{x})$ is given by the sum of the confidences of all nearby points in the computed depth-maps. Since the graph-cut algorithm **minimizes** the discontinuity cost, and we want to *maximize* the photo-consistency, $\rho(\mathbf{x})$ is simply inverted using the exponential:

$$\rho(\mathbf{x}) = e^{-\mu\mathcal{C}(\mathbf{x})}, \tag{3}$$

where μ is a very stable rate-of-decay parameter which in all our experiments was set to 0.05.

As a way of improving the big memory requirements of graph-cut methods, we propose to store the values of $\rho(\mathbf{x})$ in an octree partition of 3D space. The size of the octree voxel will depend on the photo-consistency value $\mathcal{C}(\mathbf{x})$. Voxels with a non-zero photo-consistency value will have the finest resolution while the remaining space where $\mathcal{C}(\mathbf{x}) = 0$ will be partitioned using bigger voxels, the voxel size being directly linked with the distance to the closest "non-empty" voxel (see Fig. 3 for an example of such an octree partition). As an implementation detail, the only modification needed in the graph-cut algorithm to use a multi-resolution grid is that now links between neighboring nodes need to be weighted accordingly to the surface area shared by both nodes.

3.3. Labeling cost from a set of depth-maps

In the same way as the computation of the discontinuity cost, the ballooning term $\sigma(\mathbf{x})$ can be computed exclusively from a set of depth-maps. In this paper we propose to use the probabilistic evidence for visibility introduced in Section 4 as an *intelligent* ballooning term. To do so, all we need is to choose a noise model for our sensor given a depth-map D and its confidence $\mathcal{C}(D)$. We propose to use a simplistic yet powerful model of a Gaussian contaminated with a uniform distribution, *i.e.*, an inlier model plus an outlier model. The inlier model is assumed to be a Gaussian distribution centered around the true depth. The standard deviation is considered to be a constant value that only depends on the image resolution and camera baseline. The outlier ratio varies according to the confidence on the depth

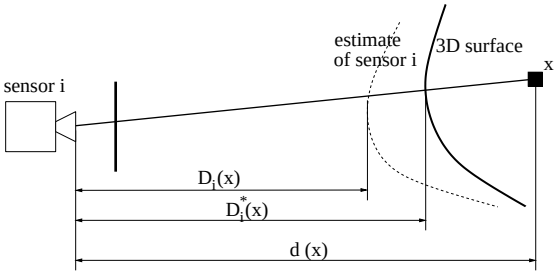

Figure 1. Sensor depth notation. Sensor i measures the depth of the scene along the optic ray from the sensor to 3D point $\mathbf{x}$. The depth of point $\mathbf{x}$ from sensor i is $d_i(\mathbf{x})$ while the correct depth of the scene along that ray is $D_i^*(\mathbf{x})$ and the sensor measurement is $D_i(\mathbf{x})$.

estimation $\mathcal{C}(D)$, and in our case is just proportional to it. The labeling cost $\sigma(\mathbf{x})$ at a given location is just the evidence of visibility. The details of this calculation are laid out in the next Section.

4. Probabilistic fusion of depth sensors

This section considers the problem of probabilistically fusing depth maps obtained from N depth sensors. We will be using the following notation: The sensor data is a set of N depth maps $\mathcal{D} = D_1, \ldots, D_N$. A 3D point $\mathbf{x}$ can therefore be projected to a pixel of the depth map of the i-th sensor and the corresponding depth measurement at that pixel is written as $D_i(\mathbf{x})$ while $D_i^*(\mathbf{x})$ denotes the true depth of the 3D scene. The measurement $D_i(\mathbf{x})$ contains some noise which is modeled probabilistically by a pdf conditional on the real surface depth

$$p\left(D_i(\mathbf{x}) \mid D_i^*(\mathbf{x})\right).$$

The depth of the point $\mathbf{x}$ away from the sensor is $d_i(\mathbf{x})$ (see figure 1). If $\mathbf{x}$ is located on the 3D scene surface then $\forall i \;\; D_i^*(\mathbf{x}) = d_i(\mathbf{x})$. If for a particular sensor i we have $D_i^*(\mathbf{x}) > d_i(\mathbf{x})$ this means that the sensor can *see beyond* $\mathbf{x}$ or in other words that $\mathbf{x}$ is *visible* from the sensor. We denote this event by $V_i(\mathbf{x})$. When the opposite event $\overline{V_i}(\mathbf{x})$ is true, as in figure 1, then $\mathbf{x}$ is said to be *occluded* from the sensor. To fuse these measurements we consider a predicate $V(\mathbf{x})$ which is read as: '$\mathbf{x}$ *is visible from at least one sensor*'. More formally the predicate is defined as follows:

$$V(\mathbf{x}) \equiv \exists i \; V_i(\mathbf{x}) \tag{4}$$

$V(\mathbf{x})$ acts as a proxy for the predicate we should ideally be examining which is '$\mathbf{x}$ *is outside the volume of the 3D scene*'. However our sensors cannot provide any evidence beyond $D_i^*(\mathbf{x})$ along the optic ray, the rest of the

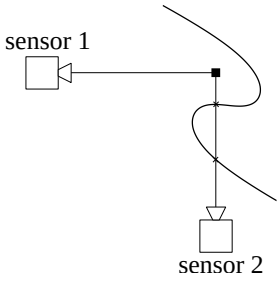

Figure 2. Visibility from sensors. In the example shown above the point is not visible from sensor 2 while it is visible from sensor 1, i.e. we have $V_1\overline{V}_2$. In the absence a surface prior that does not favor geometries such as the one shown above, one can safely assume that there is no probabilistic dependence between visibility or invisibility from any two sensors.

points on that ray being occluded. If there are locations that are occluded from all sensors, no algorithm could produce any evidence for these locations being inside or outside the volume. In that sense therefore, $V(\mathbf{x})$ is the strongest predicate one could hope for in an optical system. An intuitive assumption made throughout this paper is that the probability of $V(\mathbf{x})$ depends only on the depth measurements of sensors along optic rays that go through $\mathbf{x}$. This means that most of our inference equations will be referring to a single point $\mathbf{x}$, in which case the $\mathbf{x}$ argument can be safely removed from the predicates. Our set of assumptions which we denote by $\mathcal{J}$ consists of the following:

- The probability distributions of the true depths of the scene $D_1^*(\mathbf{x}) \cdots D_N^*(\mathbf{x})$ and also of the measurements $D_1(\mathbf{x}) \cdots D_N(\mathbf{x})$ are independent given $\mathcal{J}$ (see figure 2 for justification).
- The probability distribution of of a sensor measurement given the scene depths and all other measurements only depends on the surface depth it is measuring:

$$p\left(D_i \mid D_1^* \cdots D_N^*\, D_{j\neq i}\, \mathcal{J}\right) = p\left(D_i \mid D_i^* \mathcal{J}\right)$$

We are interested in computing the evidence function under this set of independence assumptions [14] for the visibility of the point given all the sensor measurements:

$$e\left(V \mid D_1 \cdots D_N \mathcal{J}\right) = \log \frac{p\left(V \mid D_1 \cdots D_N \mathcal{J}\right)}{p\left(\overline{V} \mid D_1 \cdots D_N \mathcal{J}\right)}.$$

From $\mathcal{J}$ and rules of probability one can derive:

$$p\left(\overline{V} \mid D_1 \cdots D_N \mathcal{J}\right) = \prod_{i=1}^{N} p\left(\overline{V}_i \mid D_i \mathcal{J}\right).$$

and

$$p\left(\overline{V}_i \mid D_i \mathcal{J}\right) = \frac{\int_0^{d_i} p\left(D_i \mid D_i^* \mathcal{J}\right) p\left(D_i^* \mid \mathcal{J}\right) dD_i^*}{\int_0^{\infty} p\left(D_i \mid D_i^* \mathcal{J}\right) p\left(D_i^* \mid \mathcal{J}\right) dD_i^*}$$

As mentioned, the distributions $p\left(D_i \mid D_i^* \mathcal{J}\right)$ encode our knowledge about the measurement model. Reasonable choices would be the Gaussian distribution or a Gaussian contaminated by an outlier process. Both of these approaches are evaluated in section 5. Another interesting option would be multi-modal distributions. The prior $p\left(D_i^* \mid \mathcal{J}\right)$ encodes some geometric knowledge about the depths in the scene. In our examples a bounding volume was given so we assumed a uniform distribution of D_i^* inside that volume.

If we write $\pi_i = p\left(\overline{V}_i \mid D_i \mathcal{J}\right)$ then the evidence for visibility is given by:

$$e\left(V \mid D_1 \cdots D_N \mathcal{J}\right) = \log \frac{1 - \pi_1 \ldots \pi_N}{\pi_1 \ldots \pi_N} \tag{5}$$

Before proceeding with the experimental evaluation of this evidence measure we point out an interesting connection between our approach and one of the classic methods in the Computer Graphics literature for merging range data.

4.1. Signed distance functions

In [7], Curless and Levoy compute signed distance functions from each depth-map (positive towards the camera and negative inside the scene) whose weighted average is then stored in a 3D scalar field. So if $w_i(\mathbf{x})$ represents the confidence of depth measurement $D_i(\mathbf{x})$ in the i-th sensor, the 3D scalar field they compute is:

$$F(\mathbf{x}) = \sum_{i=1}^{N} w_i(\mathbf{x}) \left(d_i(\mathbf{x}) - D_i(\mathbf{x})\right) \tag{6}$$

The zero level of $F(\mathbf{x})$ is then computed using marching cubes. While this method provides quite accurate results it has a drawback: For a set of depth maps around a closed object, distances from opposite sides interfere with each other. To avoid this effect [7] actually clamps the distance on either side of a depth map. The distance must be left unclamped far enough behind the depth map so that all distance functions contribute to the zero-level crossing, but not too far so as to compromise the reconstruction of thin structures. This limitation is due to the fact that the method implicitly assumes that the surface has low relief or that there are no self-occlusions. This can be expressed in several ways but perhaps the most intuitive is that every optic ray from every sensor intersects the surface only once. This means that if a point $\mathbf{x}$ is visible from at least one sensor then it must be visible from all sensors (see figure 2). Using this assumption, an analysis similar to the one in the previous section leads to some a surprising insight into the algorithm. More precisely, if we set the prior probability for visibility to $p(V) = 0.5$ and assume the logistic distribution

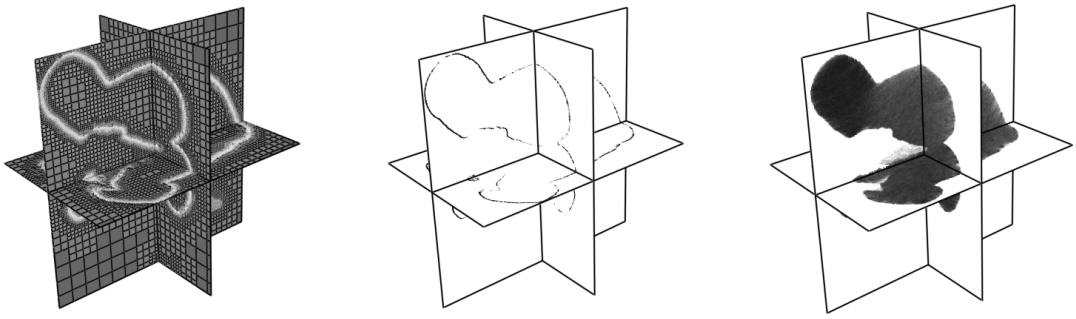

Figure 3. ***Different terms used in the graph-cut algorithm to reconstruct the Gormley sculpture.*** *Left: multi-resolution grid used in the graph-cut algorithm. Middle: Discontinuity cost $\rho(\mathbf{x})$ (or photo-consistency). Right: labeling cost $\sigma(\mathbf{x})$ (or intelligent ballooning).*

for sensor noise, i.e.

$$p\left(D_i, D_i^* \mid \mathcal{I}\right) \propto \operatorname{sech}\left(\frac{D_i^* - D_i}{2w_i}\right)^2$$

then the probabilistic evidence for V given all the data exactly corresponds to the right hand side of (6). In other words, the sum of signed distance functions of [7] can be seen as an accumulation of probabilistic evidence for visibility of points in space, given a set of noisy measurements of the depth of the 3D scene. This further reinforces the usefulness of probabilistic evidence for visibility.

5. Experiments

We present a sequence of 72 images of a "crouching man" sculpture made of plaster by the modern sculptor Antony Gormley (see Fig. 4 top). The image resolution is 5 Mpix and the camera motion was recovered by standard structure from motion techniques [22]. The object exhibits significant self-occlusions, a large concavity in the chest and two thin legs which make it a very challenging test to validate our new ballooning term. The first step in the reconstruction process is to compute a set of depth-maps from the input images. This process is by far the most expensive of the whole pipeline in terms of computation time. A single depth-map takes between 90 and 120 seconds, the overall computation time being close to 2 hours. Once the depth-maps are computed, a 3D octree grid can be built (see Fig. 3 left) together with the discontinuity cost and the labeling cost (see Fig. 3 middle and right respectively). Because of the octree grid, we are able to use up to 10 levels of resolution to compute the graph-cut, *i.e.*, the equivalent of a regular grid of 1024^3 voxels. We show in figure 4 some of the images used in the reconstruction (top), the result using an implementation of [21] (middle) and the reconstruction result of the proposed method (bottom). We can appreciate how the constant ballooning term introduced in [21] is unable to reconstruct correctly the feet and the concavities at the same time. In order to recover thin structures such as the feet, the ballooning term needs to be stronger. But even before the feet are fully recovered, the concavities start to over inflate.

Finally we show in figure 5 the effect of having an outlier component in the noise model of the depth sensor when computing the volume of evidence of visibility. The absence of an outlier model that is able to cope with noisy depth estimates appears in the volume of visibility as tunnels "drilled" by the outliers (see Fig. 5 center). Adding an outlier term clearly reduces the tunneling effect while preserving the concavities (see Fig. 5 right).

6. Conclusions

We have presented a new formulation to multi-view stereo that treats the problem as probabilistic 3D segmentation. The primary result of this investigation is that the photo-consistency criterion provides a foreground/background model that can predict if a 3D location is inside or outside the scene. This fact, which has not received much notice in the multi-view stereo literature, lets us replace the commonly used naive inflationary model with probabilistic evidence for the visibility of 3D locations. The proposed algorithm significantly outperforms ballooning based approaches, especially in concave regions and thin protrusions. We also report on a surprising connection between the proposed visibility criterion and a classic computer graphics technique for depth-map fusion, which further validates our approach. As future work we are planning a detailed evaluation of our approach with state-of-the-art depth-map fusion techniques such as [15]. We are also considering evaluating our method with the Middlebury datasets [18].

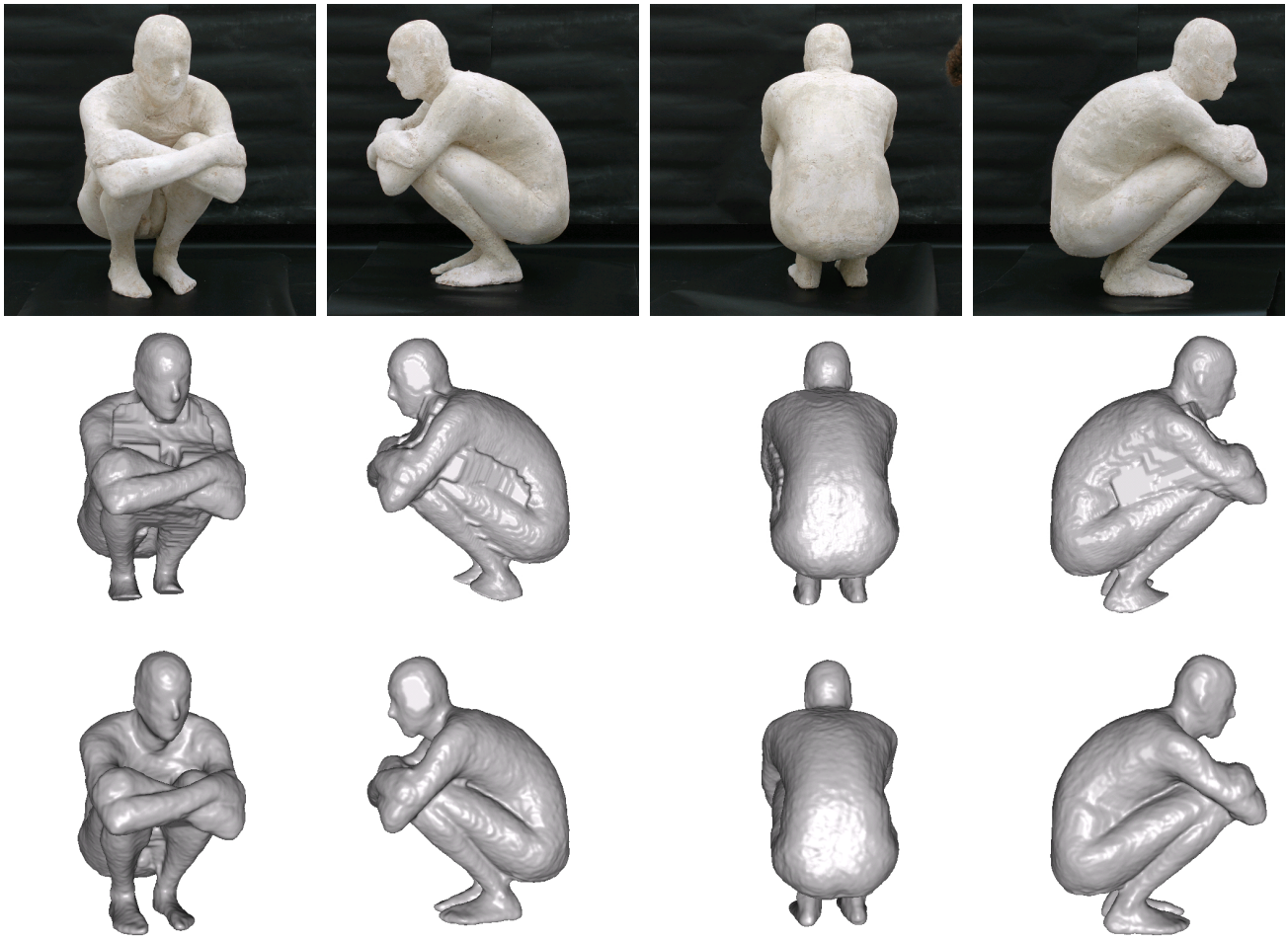

Figure 4. ***Comparison of our reconstruction results with previous methods.*** *Plaster model of a crouching man by Antony Gormley, 2006. Top: some of the input images. Middle: views of reconstructed model using the technique of [21] with a constant ballooning term. No constant ballooning factor is able to reconstruct correctly the feet and the concavities at the same time. Bottom: views of reconstructed model using the intelligent ballooning proposed in this paper and shown in Fig.5 right.*

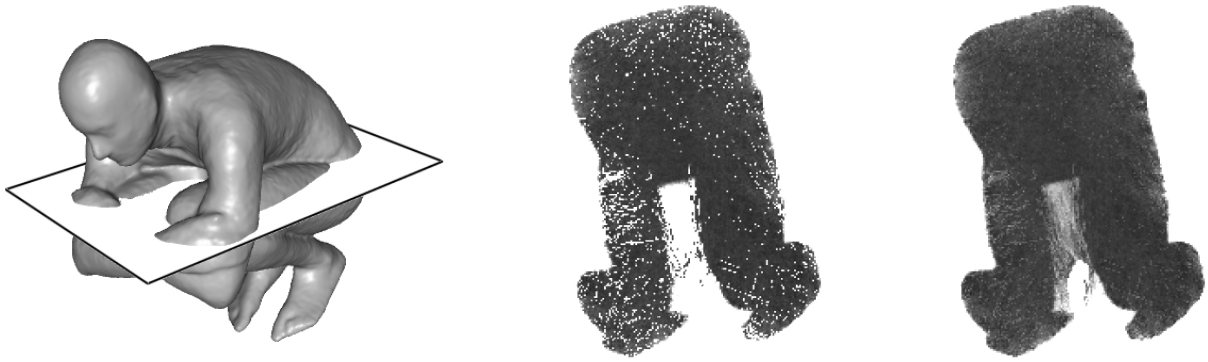

Figure 5. ***Comparison of two different inlier/outlier ratios for the depth sensor noise model.*** *Left: 3D location of one slice of the volume of "evidence of visibility". Middle: the sensor model is a pure Gaussian without any outlier model. Outliers "drill" tunnels in the visibility volume. Right: the sensor model takes into account an outlier model. The visibility volume is more robust against outliers while the concavities are still distinguishable.*

Appendix. Interpretation of signed distance functions.

Using the predicates we have already defined, the assumption of no self-occlusion can be expressed by

$$V \leftrightarrow \forall i \; V_i. \tag{7}$$

From (4) and (7) we see that if a point $\mathbf{x}$ is visible (invisible) from one sensor it is visible (invisible) from all sensors, i.e. $V_1 \leftrightarrow \cdots \leftrightarrow V_N \leftrightarrow V$. Let $\mathcal{I}$ stand for the prior knowledge which includes the geometric description of the problem and (7). Given (7) events $D_1 \cdots D_N$ are independent under the knowledge of V or $\overline{V}$ which means that using Bayes' theorem we can write:

$$p\left(V \mid D_1 \cdots D_N \mathcal{I}\right) = \frac{p\left(V \mid \mathcal{I}\right) \prod_{i=1}^{N} p\left(D_i \mid V\mathcal{I}\right)}{p\left(D_1 \cdots D_N \mid \mathcal{I}\right)} \tag{8}$$

Obtaining the equivalent equation for $\overline{V}$ and dividing with equation (8) and taking logs gives us:

$$e\left(V \mid D_1 \cdots D_N \mathcal{I}\right) = e\left(V \mid \mathcal{I}\right) + \sum_{i=1}^{N} \log \frac{p\left(D_i \mid V\mathcal{I}\right)}{p\left(D_i \mid \overline{V}\mathcal{I}\right)}. \tag{9}$$

By several applications of Bayes' theorem we get:

$$e\left(V \mid D_1 \cdots D_N \mathcal{I}\right) = \sum_{i=1}^{N} \log \frac{\alpha_i}{\beta_i} - (N-1) e\left(V \mid \mathcal{I}\right). \tag{10}$$

where $\alpha_i = \int_{d_i}^{\infty} p\left(D_i, D_i^* \mid \mathcal{I}\right) dD_i^*$ and $\beta_i = \int_0^{d_i} p\left(D_i, D_i^* \mid \mathcal{I}\right) dD_i^*$. We now set $e\left(V \mid \mathcal{I}\right) = 0$ and assume the noise model is given by the logistic function

$$p\left(D_i, D_i^* \mid \mathcal{I}\right) \propto \operatorname{sech}\left(\frac{D_i^* - D_i}{2w_i}\right)^2.$$

Using standard calculus one can obtain the following expression for the evidence

$$e\left(V \mid D_1 \cdots D_N \mathcal{I}\right) = \sum_{i=1}^{N} w_i \left(d_i - D_i\right), \tag{11}$$

equal to the average of the distance functions used in [7].□

Notation

N	Number of images/sensors
$\mathbf{x}$	3D location
$\mathcal{I}$	Prior
$\epsilon(A)$	Evidence of predicate A
$p(A)$	Probability of predicate A
$D_i(\mathbf{x})$	Depth measured by sensor i for location $\mathbf{x}$
$D_i^*(\mathbf{x})$	True depth of the scene for sensor i
$\mathcal{C}(\mathbf{x})$	Confidence of depth estimation at location $\mathbf{x}$
$V_i(\mathbf{x})$	Predicate '$\mathbf{x}$ *is visible from sensor* i'
$V(\mathbf{x})$	Predicate '$\mathbf{x}$ *is visible from at least one sensor*'

References

[1] M. Agrawal and L.-S. Davis. A probabilistic framework for surface reconstruction from multiple images. In *Proc. IEEE Conf. on CVPR*, volume 2, pages 470–477, 2001.

[2] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive GMMRF model. In *Proc. 8^{th} Europ. Conf. on Computer Vision*, pages 428–441, 2004.

[3] Y. Boykov and V. Kolmogorov. Computing geodesics and minimal surfaces via graph cuts. In *Proc. 9^{th} Intl. Conf. on Computer Vision*, pages 26–33, 2003.

[4] Y. Boykov and V. Lempitsky. From photohulls to photoflux optimization. In *Proc. BMVC, to appear*, pages 1149–1158, 2006.

[5] A. Broadhurst, T. Drummond, and R. Cipolla. A probabilistic framework for space carving. In *Proc. 8^{th} Intl. Conf. on Computer Vision*, volume 1, pages 338–393, 2001.

[6] L. Cohen and I. Cohen. Finite-element methods for active contour models and balloons for 2-d and 3-d images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(11):1131–1147, November 1993.

[7] B. Curless and M. Levoy. A volumetric method for building complex models from range images. *Proc. of the ACM SIGGRAPH*, pages 303–312, 1996.

[8] O. Faugeras and R. Keriven. Variational principles, surface evolution, pdes, level set methods and the stereo problem. *IEEE Transactions on Image Processing*, 7(3):335–344, 1998.

[9] Y. Furukawa and J. Ponce. Carved visual hulls for image-based modeling. In *Proc. 9^{th} Europ. Conf. on Computer Vision*, volume 1, pages 564–577, 2006.

[10] P. Gargallo and P. Sturm. Bayesian 3d modeling from images using multiple depth maps. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume II, pages 885–891, 2005.

[11] M. Goesele, B. Curless, and S. Seitz. Multi-view stereo revisited. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 2402–2409, 2006.

[12] C. Hernández and F. Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367–392, December 2004.

[13] A. Hornung and L. Kobbelt. Hierarchical volumetric multi-view stereo reconstruction of manifold surfaces based on dual graph embedding. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 503–510, 2006.

[14] E. Jaynes. *Probability theory: the logic of science*. Cambridge University Press, 2003.

[15] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. *Eurographics SGP*, pages 61–70, 2006.

[16] V. Lempitsky, Y. Boykov, and D. Ivanov. Oriented visibility for multiview reconstruction. In *Proc. 9^{th} Europ. Conf. on Computer Vision*, volume 3, pages 226–238, 2006.

[17] J. Pons, R. Keriven, and O. Faugeras. Modelling dynamic scenes by registering multi-view image sequences. In *Proc. IEEE Conf. on CVPR*, volume 2, pages 822–827, 2005.

[18] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 519–528, 2006.

[19] S. Sinha and M. Pollefeys. Multi-view reconstruction using photo-consistency and exact silhouette constraints: A maximum-flow formulation. In *Proc. 10^{th} Intl. Conf. on Computer Vision*, volume 1, pages 349–356, 2005.

[20] S. Tran and L. Davis. 3d surface reconstruction using graph cuts with surface constraints. In *Proc. 9^{th} Europ. Conf. on Computer Vision*, volume 2, pages 218–231, 2006.

[21] G. Vogiatzis, P. Torr, and R. Cipolla. Multi-view stereo via volumetric graph-cuts. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 391–398, 2005.

[22] A. Zisserman and R. Hartley. *Multiple View Geometry*. Springer-Verlag, 2000.