to appear in IEEE Conf. on Computer Vision and Pattern Recognition'07

3D Occlusion Inference from Silhouette Cues

Li Guan, Jean-Sébastien Franco and Marc Pollefeys
UNC Chapel Hill, USA
lguan@cs.unc.edu, franco@cs.unc.edu, marc@cs.unc.edu

Abstract

We consider the problem of detecting and accounting for the presence of occluders in a 3D scene based on silhouette cues in video streams obtained from multiple, calibrated views. While well studied and robust in controlled environments, silhouette-based reconstruction of dynamic objects fails in general environments where uncontrolled occlusions are commonplace, due to inherent silhouette corruption by occluders. We show that occluders in the interaction space of dynamic objects can be detected and their 3D shape fully recovered as a byproduct of shape-from-silhouette analysis. We provide a Bayesian sensor fusion formulation to process all occlusion cues occurring in a multi-view sequence. Results show that the shape of static occluders can be robustly recovered from pure dynamic object motion, and that this information can be used for online self-correction and consolidation of dynamic object shape reconstruction.

1. Introduction

Silhouette-based approaches are popular in multi-view environments, as they are relevant to many computer vision problems such as 3D modeling, object detection, recognition, tracking, and are useful for a very wide range of applications such as 3D photography, virtual reality and real-time human-computer interaction. Such techniques have been shown particularly robust and successful in controlled, unoccluded environments. A major challenge is to apply silhouette-based approaches in uncontrolled, outdoor environments where occlusions are common and unavoidable.

In this paper, we show that the shape of static occluders in the interaction space of moving objects can be recovered online by accumulating occlusion cues from dynamic object motion, using a Bayesian sensor fusion framework. Moving objects and static occluders can then be estimated cooperatively in an online process. The type of solutions proposed here are immediately relevant to silhouette-based techniques, but can also be useful for a wider range of multi-view vision problems where such occlusion detection can yield benefits.

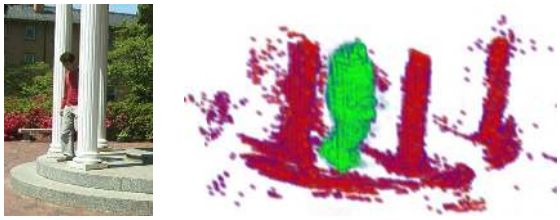

Figure 1. Overview. (left) One of 9 observed views of a scene, where a person walks occluded by pillars. (right) Occluder and person shape retrieved cooperatively using our approach.

A popular approach to model the shapes of objects is to learn the appearance of background images [3, 15], perform automatic silhouette extraction, and use Shape-from-Silhouette techniques (SfS) [2, 13, 16]. However occlusions with background objects whose appearance is also recorded in background images have a negative impact over silhouette-based modeling, because extracted silhouettes can become incomplete. In particular the inclusive property of visual hulls [11] with respect to the object it models is no longer guaranteed. This limitation comes on top of the usual sensitivities to noise, shadows, and lighting conditions. Moving the problem outdoors greatly increases the difficulty of dealing with silhouettes for all of the above reasons. Improving the robustness of silhouette methods has been explored, by using silhouette priors of multi-view sets [9], spatio-temporal regularization [8], silhouette cue integration using a sensor fusion paradigm [6] or a discrete optimization scheme [7, 14]. While decreasing the various sensitivities and failures arising from partial occlusions and silhouette corruption, these works do not address explicit detection of 3D occluders as proposed.

More generally detecting and accounting for occlusions has attracted the attention of researchers for problems such as structure from motion [5], motion and occlusion boundary detection [1]. The scope of these works is however limited to extraction of sparse 2D features such as T-junctions or edges to improve robustness of data estimation. A recent method proposes a 2D solution in the form of occluder maps and account for them in building the visual hull of dynamic objects [10]. To the best of our knowledge no method has addressed the dense recovery of occluder shapes from

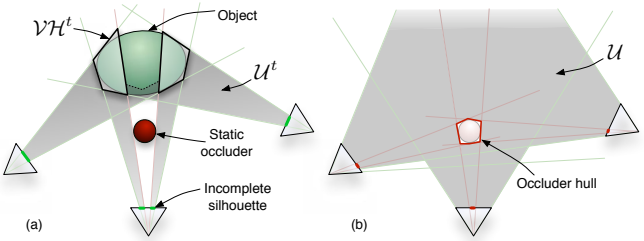

Figure 2. Deterministic occlusion reasoning. By building a visual hull based on incomplete silhouettes, we can deduce a conservative free-space region (dark gray), free of static occluders.

multiple views as proposed in the present work.

1.1. Principle

We examine video sequences obtained from n fully calibrated cameras, observing a scene at discrete time steps $t \in \{1, \cdots, T\}$ where people, and more generally dynamic, moving objects can evolve. A set of *background images* of the scene, free from any *dynamic object*, have previously been observed for each camera. *Static occluder objects*, whose appearance is recorded in the background images of the scene, are present in the interaction space of dynamic objects. They are thus liable to generate partial occlusions of dynamic objects, with respect to one or several cameras.

Theoretically occluder shapes can be accessed by carefully reasoning on the visual hull of incomplete silhouettes (Fig. 2). Let $\mathcal{S}^t$ be the set of incomplete silhouettes obtained at time t, and $\mathcal{VH}^t$ the incomplete visual hull obtained using these silhouettes. These entities are said to be incomplete because the silhouettes used are potentially corrupt by static occluders that mask the silhouette extraction process. However the incomplete visual hull is a region that is observed by all cameras as being both occupied by an object and unoccluded from any view. Thus we can deduce an entire region $\mathcal{U}^t$ of points in space that are free from any static occluder shape. $\mathcal{U}^t$ is the set of points $X \in \mathbb{R}^3$ for which a view i exists, such that the viewing line of X from view i hits the incomplete visual hull at a first visible point A_i, and $X \in O_i A_i$, with O_i the optical center of view i (Fig. 2(a)). The latter expresses the condition that X appears in front of the visual hull with respect to view i. The region $\mathcal{U}^t$ varies with t, thus assuming static occluders and broad coverage of the scene by dynamic object motion, the free space in the scene can be deduced as the region $\mathcal{U} = \bigcup_{t=1}^{T} \mathcal{U}^t$. The shape of occluders, including concavities if they were covered by object motion, can be recovered as the complement of $\mathcal{U}$ in the common visibility region of all views (Fig. 2(b)).

However this deterministic approach would yield an impractical and non-robust solution, due to inherent silhouette extraction sensitivities to noise and corruption that contribute irreversibly to the result. It also suffers from the limitation that only portions of objects that are seen by all views can contribute to occlusion reasoning. Also, this scheme only accumulates *negative* information, where occluders are certain not to be. However *positive* information is also underlying to the problem: had we known or taken a good guess at where the object shape was (which current shape-from-silhouette methods are able to provide), discrepancies between the object's projection and the actual silhouette recorded would tell us where an occlusion is positively happening. To lift these limitations and provide a robust solution, we propose a probabilistic approach to occlusion reasoning, in which all negative and positive cues are fused and compete in a complementary way toward occluder shape estimation.

1.2. Probabilistic Approach

In this paper we formulate occlusion inference as the separate Bayesian estimation, for each voxel in a 3D grid sampling the acquisition space, of how likely it is occupied by a static occluder object. We use the sensor fusion paradigm, and borrow simplifications and semantics from the occupancy grid framework of robotics [4, 12]. By modeling the likely responses in images to a known state of the scene through a generative *sensor model*, the strategy is then to use Bayes' Rule in order to solve the inverse problem, and find the likely state of occluder occupancy given noisy image data.

However, both the foreground and occluder object shapes are unknowns in this setup. While it is theoretically possible to model the problem as a joint estimation of these two shapes, this would translate in a global optimization problem over the conjunction of both shape spaces, because estimation of a voxel's state bares dependencies with all other voxels on its viewing lines with respect to all cameras. To benefit from the locality that makes occupancy grid approaches practical and efficient, we break the estimation into two steps: first estimating the occupancy of the visual hull of dynamic object's from silhouette information, using earlier work from Franco & Boyer [6] robust to partial occlusions, then estimate per-voxel occluder occupancy in a second inference, using the result of the first estimation as prior of dynamic object occupancy.

We first describe how this idea translates into a tractable Bayesian formulation, by establishing the problem's meaningful variables (§2), considering the necessary dependencies between them and decomposing their joint probability distribution (§3). We then assign parametric forms to each decomposition term that describe how silhouettes are formed given the state of the scene. The occluder occupancy at a given voxel can then be inferred using this model (§3.3). We examine a set of extensions of the method as applied to the online estimation of occluder shapes (§4.1), and refinement of dynamic objects estimation (§4.2).

2. Modeling

Consider a scene observed by n calibrated cameras. We focus on the case of one scene voxel with 3D position X among the possible coordinates in the lattice chosen for scene discretization. The two possible states of occluder occupancy at this voxel are expressed using a binary variable $\mathcal{O}$. This state is assumed to be fixed over the entire experiment in this setup under the assumption that the occluder is static. Clearly, the regions of importance to infer $\mathcal{O}$ are the n viewing lines $\mathcal{L}_i$, $i \in \{1, \cdots, n\}$, as shown in Fig. 3(a). Scene states are observed for a finite number of time instants $t \in \{1, \cdots, T\}$. In particular, dynamic visual hull occupancies of voxel X at time t are expressed by a binary statistical variable $\mathcal{G}^t$, treated as an unobserved variable to retain the probabilistic information given by [6].

2.1. Observed Variables

The voxel X projects to n image pixels $x_i, i \in 1, \cdots, n$, whose color observed at time t in view i is expressed by the variable $\mathcal{I}_i^t$. We assume that static background images were observed free of dynamic objects, and that the appearance and variability of background colors for pixels x_i was recorded and modeled using a set of parameters $\mathcal{B}_i$. Such observations can be used to infer the probability of dynamic object occupancy in the absence of background occluders. The problem of recovering occluder occupancy is more complex because it requires modeling interactions between voxels on the same viewing lines. Relevant statistical variables are shown in Fig. 3(b).

2.2. Viewing Line Modeling

Because of potential mutual occlusions, one must account for other occupancies along the viewing lines of X to infer $\mathcal{O}$. These can be either other static occluder states, or dynamic object occupancies which vary across time. Several such occluders or objects can be present along a viewing line, leading to a number of possible occupancy states for voxels on the viewing line of X. Accounting for the combinatorial number of possibilities for voxel states along X's viewing line is neither necessary nor meaningful: first because occupancies of neighboring voxels are fundamentally correlated to the presence or the absence of a single common object, second because the main useful information one needs to know to make occlusion decisions about X is to know whether something is in front of it or behind it, regardless of where along the viewing line.

With this in mind, we model each viewing line using three components, that model the state of X, the state of occlusion of X by anything in front, and the state of what is at the back of X. We model the front and back components by extracting the two most influential modes in front and behind of X, that are given by two voxels $\hat{X}_i^t$ and $\check{X}_i^t$. We select $\hat{X}_i^t$ as the voxel at time t that most contributes to the belief that X is obstructed by a dynamic object along $\mathcal{L}_i$, and $\check{X}_i^t$ as the voxel most likely to be occupied by a dynamic object behind X on $\mathcal{L}_i$ at time t.

2.3. Viewing Line Unobserved Variables

With this three component modeling, comes a number of related statistical variables illustrated in Fig. 3(b). The occupancy of voxels $\hat{X}_i^t$ and $\check{X}_i^t$ by the visual hull of a dynamic object at time t on $\mathcal{L}_i$ is expressed by two binary state variables, respectively $\hat{\mathcal{G}}_i^t$ and $\check{\mathcal{G}}_i^t$. Two binary state variables $\hat{\mathcal{O}}_i^t$ and $\check{\mathcal{O}}_i^t$ express the presence or absence of an occluder at voxels $\hat{X}_i^t$ and $\check{X}_i^t$ respectively. Note the difference in semantics between the two variable groups $\hat{\mathcal{G}}_i^t, \check{\mathcal{G}}_i^t$ and $\hat{\mathcal{O}}_i^t, \check{\mathcal{O}}_i^t$. The former designates dynamic visual hull occupancies of different time instants and chosen positions, while the latter expresses *static* occluder occupancies, whose *position only* was chosen in relation to t. Both need to be considered because they both influence the occupancy inference and are not independent. For legibility, we occasionally refer to the conjunction of a group of variables by dropping indices and exponents, *e.g.* $\mathcal{G} = \{\mathcal{G}^1, \cdots, \mathcal{G}^T\}, \mathcal{B} = \{\mathcal{B}_1, \cdots, \mathcal{B}_n\}$.

3. Joint Distribution

As a further step toward offering a tractable solution to occlusion occupancy inference, we describe the noisy interactions between the variables considered, through the decomposition of their joint probability distribution $p(\mathcal{O}, \mathcal{G}, \hat{\mathcal{O}}, \hat{\mathcal{G}}, \check{\mathcal{O}}, \check{\mathcal{G}}, \mathcal{I}, \mathcal{B})$. We propose the following:

$$p(\mathcal{O}) \prod_{t=1}^{T} p(\mathcal{G}^t|\mathcal{O}) \prod_{i=1}^{n} p(\hat{\mathcal{O}}_i^t) p(\hat{\mathcal{G}}_i^t|\hat{\mathcal{O}}_i^t) p(\check{\mathcal{O}}_i^t) p(\check{\mathcal{G}}_i^t|\check{\mathcal{O}}_i^t) \quad (1)$$
$$p(\mathcal{I}_i^t|\hat{\mathcal{O}}_i^t, \hat{\mathcal{G}}_i^t, \mathcal{O}, \mathcal{G}^t, \check{\mathcal{O}}_i^t, \check{\mathcal{G}}_i^t, \mathcal{B}_i).$$

$p(\mathcal{O})$, $p(\hat{\mathcal{O}}_i^t)$, and $p(\check{\mathcal{O}}_i^t)$ are priors of occluder occupancy. We set them to a single constant distribution $\mathcal{P}_o$ which reflects the expected ratio between occluder and non-occluder voxels in a scene. No particular region of space is to be favored *a priori*.

3.1. Dynamic Occupancy Priors

$p(\mathcal{G}^t|\mathcal{O})$, $p(\hat{\mathcal{G}}_i^t|\hat{\mathcal{O}}_i^t)$, $p(\check{\mathcal{G}}_i^t|\check{\mathcal{O}}_i^t)$ are priors of dynamic visual hull occupancy with identical semantics. This choice of terms reflects the following modeling decisions. First, the dynamic visual hull occupancies involved are considered independent of one another as they synthesize the information of three distinct regions for each viewing line. However they depend upon the knowledge of occluder occupancy at the corresponding voxel position, because occluder and dynamic object occupancies are mutually exclusive at a given scene location. Importantly however, we do not have direct

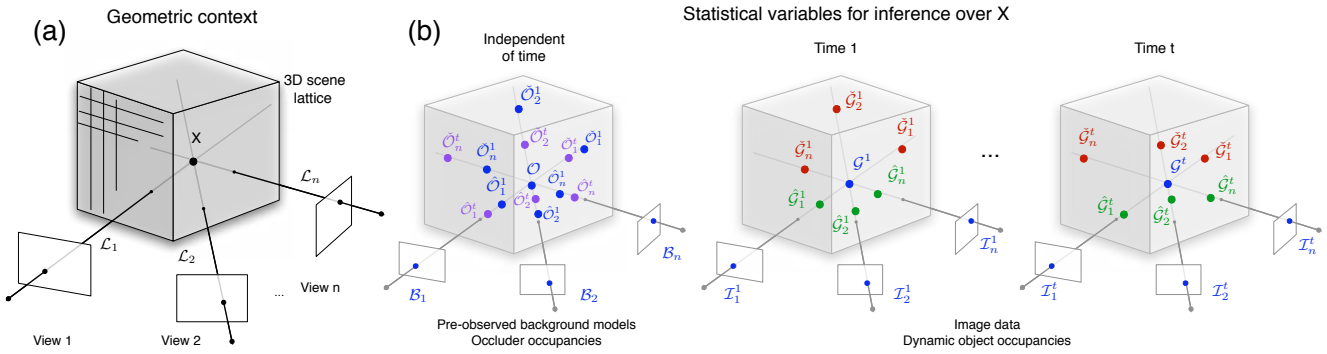

Figure 3. Problem overview. Statistical variables and their geometric/temporal meaning.

access to dynamic object occupancies but to the occupancies of its *visual hull*. Fortunately this ambiguity can be adequately modeled in a Bayesian framework, by introducing a local hidden variable $\mathcal{C}$ expressing the correlation between dynamic and occluder occupancy:

$$p(\mathcal{G}^t|\mathcal{O}) = \sum_{\mathcal{C}} p(\mathcal{C})p(\mathcal{G}^t|\mathcal{C},\mathcal{O}). \tag{2}$$

We set $p(\mathcal{C}=1) = \mathcal{P}_c$ using a constant expressing our prior belief about the correlation between visual hull and occluder occupancy. The prior $p(\mathcal{G}^t|\mathcal{C},\mathcal{O})$ explains what we expect to know about $\mathcal{G}^t$ given the state of $\mathcal{C}$ and $\mathcal{O}$:

$$p(\mathcal{G}^t=1|\mathcal{C}=0,\mathcal{O}=\omega) = \mathcal{P}_{\mathcal{G}^t} \quad \forall\omega \tag{3}$$
$$p(\mathcal{G}^t=1|\mathcal{C}=1,\mathcal{O}=0) = \mathcal{P}_{\mathcal{G}^t} \tag{4}$$
$$p(\mathcal{G}^t=1|\mathcal{C}=1,\mathcal{O}=1) = \mathcal{P}_{go}, \tag{5}$$

with $\mathcal{P}_{\mathcal{G}^t}$ the prior dynamic object occupancy probability as computed independently of occlusions [6], and $\mathcal{P}_{go}$ set close to 0, expressing that it is unlikely that the voxel is occupied by dynamic object visual hulls when the voxel is known to be occupied by an occluder and both dynamic and occluder occupancy are known to be strongly correlated (5). The probability of visual hull occupancy is given by the previously computed occupancy prior, in case of non-correlation (3), or when the states are correlated but occluder occupancy is known to be empty (4).

3.2. Image Sensor Model

The sensor model $p(\mathcal{I}_i^t|\hat{\mathcal{O}}_i^t,\hat{\mathcal{G}}_i^t,\mathcal{O},\mathcal{G}^t,\check{\mathcal{O}}_i^t,\check{\mathcal{G}}_i^t,\mathcal{B}_i)$ is governed by a hidden local per-pixel process $\mathcal{S}$. The binary variable $\mathcal{S}$ represents the hidden silhouette detection state (0 or 1) at this pixel. It is unobserved information and can be marginalized, given an adequate split into two subterms:

$$\begin{aligned} &p(\mathcal{I}_i^t|\hat{\mathcal{O}}_i^t,\hat{\mathcal{G}}_i^t,\mathcal{O},\mathcal{G}^t,\check{\mathcal{O}}_i^t,\check{\mathcal{G}}_i^t,\mathcal{B}_i) \\ &= \sum_{\mathcal{S}} p(\mathcal{I}_i^t|\mathcal{S},\mathcal{B}_i)p(\mathcal{S}|\hat{\mathcal{O}}_i^t,\hat{\mathcal{G}}_i^t,\mathcal{O},\mathcal{G}^t,\check{\mathcal{O}}_i^t,\check{\mathcal{G}}_i^t). \end{aligned} \tag{6}$$

$p(\mathcal{I}_i^t|\mathcal{S},\mathcal{B}_i)$ indicates what color distribution we expect to observe given the knowledge of silhouette detection and background color model at this pixel. When $\mathcal{S}=0$, the silhouette is undetected and thus the color distribution is dictated by the pre-observed background model $\mathcal{B}_i$ (considered Gaussian in our experiments). When $\mathcal{S}=1$, a dynamic object's silhouette is detected, in which case our knowledge of color is limited, thus we use a uniform distribution in this case, favoring no dynamic object color *a priori*.

$p(\mathcal{S}|\hat{\mathcal{O}}_i^t,\hat{\mathcal{G}}_i^t,\mathcal{O},\mathcal{G}^t,\check{\mathcal{O}}_i^t,\check{\mathcal{G}}_i^t)$ is the second part of the sensor model, which explicits what silhouette state is expected to be observed given the three dominant occupancy state variables of the corresponding viewing line. Since these are encountered in the order of visibility $\hat{X}_i^t$, X, $\check{X}_i^t$, the following relations hold:

$$\begin{aligned} &p(\mathcal{S}|\{\hat{\mathcal{O}}_i^t,\hat{\mathcal{G}}_i^t,\mathcal{O},\mathcal{G}^t,\check{\mathcal{O}}_i^t,\check{\mathcal{G}}_i^t\}=\{o,g,\mathrm{k},\mathrm{l},\mathrm{m},\mathrm{n}\},\mathcal{B}_i) \\ &=p(\mathcal{S}|\{\hat{\mathcal{O}}_i^t,\hat{\mathcal{G}}_i^t,\mathcal{O},\mathcal{G}^t,\check{\mathcal{O}}_i^t,\check{\mathcal{G}}_i^t\}=\{0,0,o,g,\mathrm{p},\mathrm{q}\},\mathcal{B}_i) \\ &=p(\mathcal{S}|\{\hat{\mathcal{O}}_i^t,\hat{\mathcal{G}}_i^t,\mathcal{O},\mathcal{G}^t,\check{\mathcal{O}}_i^t,\check{\mathcal{G}}_i^t\}=\{0,0,0,0,o,g\},\mathcal{B}_i) \\ &=P_S(\mathcal{S}|o,g) \qquad \forall(o,g)\neq(0,0)\ \ \forall(\mathrm{k},\mathrm{l},\mathrm{m},\mathrm{n},\mathrm{p},\mathrm{q}). \end{aligned} \tag{7}$$

These expressions convey two characteristics. First, that the form of this distribution is given by the first non-empty occupancy component in the order of visibility, regardless of what is behind this component on the viewing line. Second, that the form of the first non-empty component is given by an identical sensor prior $P_S(\mathcal{S}|o,g)$. We set the four parametric distributions of $P_S(\mathcal{S}|o,g)$ as following:

$$P_S(\mathcal{S}=1|0,0)=\mathcal{P}_{fa} \qquad P_S(\mathcal{S}=1|1,0)=\mathcal{P}_{fa} \tag{8}$$
$$P_S(\mathcal{S}=1|0,1)=\mathcal{P}_{d} \qquad P_S(\mathcal{S}=1|1,1)=0.5, \tag{9}$$

where $\mathcal{P}_{fa}\in[0,1]$ and $\mathcal{P}_d\in[0,1]$ are constants expressing the prior probability of *false alarm* and the probability of *detection*, respectively. They can be chosen once for all datasets as the method is not sensitive to the exact value of these priors. Meaningful values for $\mathcal{P}_{fa}$ are close to 0, while $\mathcal{P}_d$ is generally close to 1. (8) expresses the cases where no silhouette is expected to be detected in images, i.e. either when there are no objects at all on the viewing line,

or when the first encountered object is a static occluder, respectively. (9) expresses two distinct cases. First, the case where a dynamic object's visual hull is encountered on the viewing line, in which case we expect to detect a silhouette at the matching pixel. Second, the case where both an occluder and dynamic visual hull are present at the first non-free voxel. This is perfectly possible, because the visual hull is an overestimate of the true dynamic object shape. While the true shape of objects and occluders are naturally mutually exclusive, the *visual hull* of dynamic objects can overlap with occluder voxels. In this case we set the distribution to uniform, because the silhouette detection state cannot be predicted: it can be caused by shadows casted by dynamic objects on occluders in the scene, and noise.

3.3. Inference

Estimating the occluder occupancy at a voxel translates to estimating $p(\mathcal{O}|\mathcal{IB})$ in Bayesian terms. Applying Bayes rule to the modeled joint probability (1) leads to the following expression, once hidden variable marginalizations are appropriately factorized:

$$p(\mathcal{O}|\mathcal{IB}) = \frac{1}{z}\, p(\mathcal{O}) \prod_{t=1}^{T} \left(\sum_{\mathcal{G}^t} p(\mathcal{G}^t|\mathcal{O}) \left(\prod_{i=1}^{n} \mathcal{P}_i^t \right) \right) \quad (10)$$

$$\text{where } \mathcal{P}_i^t = \sum_{\check{\mathcal{O}}_i^t, \check{\mathcal{G}}_i^t} p(\check{\mathcal{O}}_i^t) p(\check{\mathcal{G}}_i^t|\check{\mathcal{O}}_i^t) \sum_{\hat{\mathcal{O}}_i^t, \hat{\mathcal{G}}_i^t} p(\hat{\mathcal{O}}_i^t) p(\hat{\mathcal{G}}_i^t|\hat{\mathcal{O}}_i^t)$$
$$p(\mathcal{I}_i^t|\hat{\mathcal{O}}_i^t, \hat{\mathcal{G}}_i^t, \mathcal{O}, \mathcal{G}^t, \check{\mathcal{O}}_i^t, \check{\mathcal{G}}_i^t, \mathcal{B}_i). \quad (11)$$

$\mathcal{P}_i^t$ expresses the contribution of view i at a time t. The formulation therefore expresses Bayesian fusion over the various observed time instants and available views, with marginalization over unknown viewing line states (10). The normalization constant z is easily obtained by ensuring summation to 1 of the distribution.

4. Extensions and Applications

The proposed formulation performs a fusion of occlusion cues across different views and time instants indifferently of the order. This enables incremental updates of the inference, and opens the prospect of online applications such as cooperative estimation of occlusion and dynamic object shape, which we examine in §4.2. However an occluder estimation bias can occur in partially observed regions of space, because the formulation implicitly assumes that occlusion cues were well sampled by dynamic object motion within the scene. It is thus necessary to determine when enough occlusion cues have been accumulated for the estimation to be reliable, in an online processing framework.

4.1. Online Incremental Computation

To determine the reliability of voxels, we model the intuition that voxels whose occlusion cues arise from an abnormally low number of views should not be trusted. Since this clause involves all cameras and their observations jointly, the inclusion of this constraint in our initial model would break the symmetry in the inference formulated in (10) and defeat the possibility for online updates. Instead, we opt to use a second criterion in the form of a reliability measure $R \in [0, 1]$. Small values indicate poor coverage of dynamic objects, while large values indicate sufficient cue accumulation. We define reliability using the following expression:

$$R = \frac{1}{n} \sum_{i=1}^{n} \max_t \left(1 - \mathcal{P}_{\hat{\mathcal{G}}_i^t}\right) \mathcal{P}_{\check{\mathcal{G}}_i^t} \quad (12)$$

with $\mathcal{P}_{\hat{\mathcal{G}}_i^t}$ and $\mathcal{P}_{\check{\mathcal{G}}_i^t}$ the prior probabilities of dynamic visual hull occupancy. R examines, for each camera i, the maximum occurrence across the examined time sequence of X to be both unobstructed and in front of a dynamic object. This determines how well a given view i was able to contribute to the estimation across the sequence. R then averages these values across views, to measure the overall quality of observation, and underlying coverage of dynamic object motion for the purpose of occlusion inference.

The reliability R can be used online in conjunction to the occlusion probability estimation to evaluate a conservative occluder shape at all times, by only considering voxels for which R exceeds a certain quality threshold. As shown in §5, it can be used to reduce the sensitivity to noise in regions of space that have only been observed marginally.

4.2. Accounting for Occlusion in SfS

As more data becomes available and reliable, the results of occluder estimation can be accounted for when inferring the occupancies of dynamic objects. This translates to the evaluation of $p(\mathcal{G}^\tau|\mathcal{I}^\tau\mathcal{B})$ for a given voxel X and time τ. The difference with the classical single-frame formulation of dynamic object occupancy [6] is that we now have a prior over the occlusions at every voxel in the grid. For this inference $\mathcal{G}^\tau$ is considered independent of $\mathcal{G}^t$ $\forall t \neq \tau$, leading to the following simplified joint probability distribution:

$$p(\mathcal{O})p(\mathcal{G}^\tau|\mathcal{O}) \prod_{i=1}^{n} p(\hat{\mathcal{O}}_i^\tau) p(\hat{\mathcal{G}}_i^\tau|\hat{\mathcal{O}}_i^\tau) p(\mathcal{I}_i^\tau|\hat{\mathcal{O}}_i^\tau, \hat{\mathcal{G}}_i^\tau, \mathcal{O}, \mathcal{G}^\tau, \mathcal{B}_i),$$

where $\mathcal{G}^\tau$ and $\mathcal{O}$ are the dynamic and occluder occupancy at the inferred voxel, $\hat{\mathcal{O}}_i^\tau$, $\hat{\mathcal{G}}_i^\tau$ the variables matching the most influential component along $\mathcal{L}_i$, *in front of* X. This component is selected as the voxel whose prior of being occupied is maximal, as computed to date by occlusion inference. In this inference, there is no need to consider voxels behind X, because knowledge about their occlusion occupancy has no influence on the state of X.

The parametric forms of this distribution have identical semantics as previously but different assignments because

of the nature of the inference. Naturally no prior information about dynamic occupancy is assumed here. $p(\mathcal{O})$ and $p(\hat{\mathcal{O}}_i^\tau)$ are set using the result to date of expression (10) at their respective voxels, as prior. $p(\mathcal{G}^\tau|\mathcal{O})$ and $p(\hat{\mathcal{G}}_i^\tau|\hat{\mathcal{O}}_i^\tau)$ are constant: $p(\mathcal{G}^\tau{=}1|\mathcal{O}{=}0){=}0.5$ expresses a uniform prior for dynamic objects when the voxel is known to be occluder free. $p(\mathcal{G}^\tau{=}1|\mathcal{O}{=}1){=}\mathcal{P}_{go}$ expresses a low prior of dynamic visual hull occupancy given the knowledge of occluder occupancy, as in (5). The term $p(\mathcal{I}_i^\tau|\hat{\mathcal{O}}_i^\tau, \hat{\mathcal{G}}_i^\tau, \mathcal{O}, \mathcal{G}^\tau, \mathcal{B}_i)$ is set identically to expression (7), only stripped of the influence of $\check{\mathcal{O}}_i^\tau$, $\check{\mathcal{G}}_i^\tau$.

5. Results

5.1. Sensor Model Summary

The core of the occlusion formulation we present is controlled by five parameters $\mathcal{P}_o$, $\mathcal{P}_{go}$, $\mathcal{P}_c$, $\mathcal{P}_d$ and $\mathcal{P}_{fa}$. If two dynamic objects are *perfectly* known to occupy space in regions Ω_1 and Ω_2 (Fig. 4), various regions of importance appear in the occlusion inference, for a given camera and time instant. $\mathcal{N}_1$ and $\mathcal{N}_2$ are regions where the current view does not contribute and the inference reverts to the prior $\mathcal{P}_o$: $\mathcal{N}_1$ because it is outside of the viewing cone of dynamic objects, $\mathcal{N}_2$ because it is obstructed by an actual dynamic object Ω_1. $\mathcal{E}$ projects to a positive silhouette response area in the image and the probability of occluder occupancy is thus deduced to be low. $\mathcal{D}$ projects to an image area with low silhouette response, despite being in front of Ω_1, thus it is deduced that an occluder is probably in this region. The strength of the contribution in these regions depends on our confidence in observations, as expressed by $\mathcal{P}_d$ and $\mathcal{P}_{fa}$. Finally, Ω_1 and Ω_2 also contribute directly to the estimation through $\mathcal{P}_c$ and $\mathcal{P}_{go}$: a higher $\mathcal{P}_c$ and lower $\mathcal{P}_{go}$ give more weight to the mutual exclusivity constraint between occluders and dynamic objects and thus leads to lower occluder probabilities in these regions.

Depending on the actual probabilities of silhouette response and on the prior probabilities of the dynamic occupancy in regions Ω_1 and Ω_2, actual voxel contributions exhibit a mixture of these different behaviors in practice, which the model automatically integrates. Values given to the model parameters could be learned using training data. Nevertheless the inference has low sensitivity to small changes of these parameters, and they are sufficiently

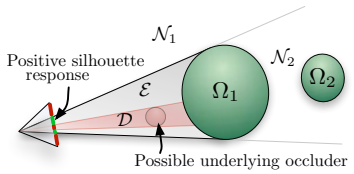

Figure 4. Regions of influence of a mono-camera sensor among the various voxels of a scene, as described by our model.

generic and intuitive that setting them manually for a large number of different sequences is possible. We use a single set of parameters throughout our experiments: $\mathcal{P}_o = 0.15$, $\mathcal{P}_{go} = 0.001$, $\mathcal{P}_c = 0.5$, $\mathcal{P}_d = 0.8$ and $\mathcal{P}_{fa} = 0.1$.

5.2. Occlusion Inference Results

We have performed several multi-view acquisitions for the purpose of validation, yielding three sequences: the PILLARS and SCULPTURE sequences which are acquired outdoors, and the CHAIR sequence, acquired indoors, with combined artificial and natural light from large bay windows. In all sequences 9 DV cameras surround the scene of interest, background models are learned in the absence of moving objects. One or several people then walk around and through the occluder in each scene. The shape of the people is estimated at each considered time step and used as prior to occlusion inference. The data is used to compute an estimate of the occluder's shape using (10). Results are presented in Fig. 5 and in the supplemental video.

The 9 cameras are geometrically calibrated, using Bouguet's toolbox based on [17], and recording at 30Hz. Color calibration is unnecessary because the model uses silhouette information only. The background model is learned per-view using a single Gaussian color model per pixel, and training images. Although simple, the model proves sufficient, even in outdoor sequences subject to background motion, foreground object shadows, and substantial illumination changes, illustrating the strong robustness of the method to difficult real conditions. The method can cope well with background misclassifications that do not lead to large coherent false positive dynamic object estimations: pedestrians are routinely seen in the background for the SCULPTURE and PILLARS sequences (e.g. Fig. 5(a1)), without any significant corruption of the inference.

Adjacent frames in the input videos contain largely redundant information for occluder modeling, thus videos can safely be subsampled. PILLARS was processed using 50% of the frames (1053 frames processed), SCULPTURE and CHAIR with 10% (160 and 168 processed frames respectively). Processing of both dynamic and occluder occupancy was handled on a 2.8 GHz PC at approximately 1 timestep per minute. The very strong locality inherent to the algorithm and preliminary benchmarks suggest that real-time performance could be achieved using a GPU implementation. Occluder information does not need to be processed for every frame because of adjacent frame redundancy, opening the possibility for online, asynchronous cooperative computation of occluder and dynamic objects at interactive frame rates.

5.3. Online Computation Results

All experiments can be computed using incremental inference updates. Fig. 6 depicts the inference's progression,

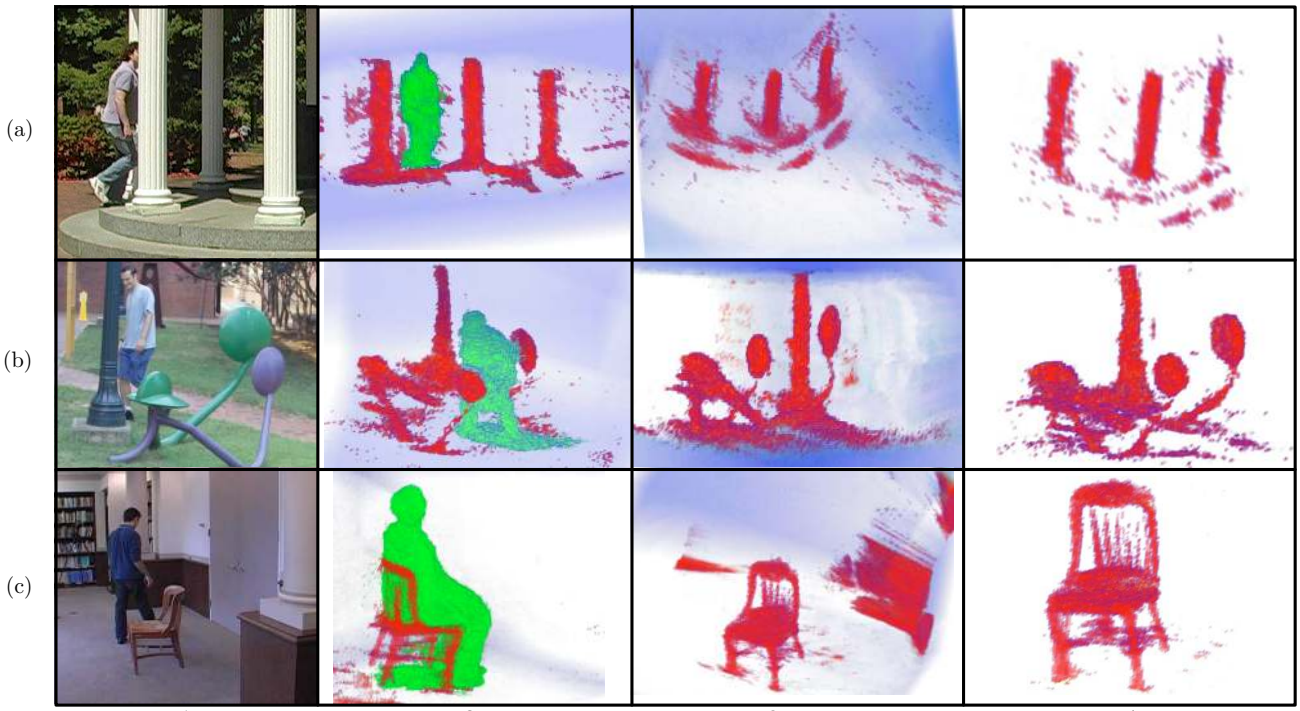

Figure 5. Occluder shape retrieval results (**best viewed in color**). Sequences: (a) PILLARS, (b) SCULPTURE, (c) CHAIR. 1) Scene overview. Note the harsh light, difficult backgrounds for (a) and (b), and specularity of the sculpture, causing no significant modeling failure. 2-3) Occluder inference according to (10). Blue: neutral regions (prior $\mathcal{P}_o$), red: high probability regions. Brighter/clear regions indicate the inferred absence of occluders. Fine levels of detail are modeled, sometimes lost - mostly to calibration. In (a) the structure's steps are also detected. 4) Same inference with additional exclusion of zones with reliability under 0.8. Peripheral noise and marginally observed regions are eliminated. The background protruding shape in (c3) is due to a single occlusion from view (c1). The supplemental video shows extensive results with these datasets, including one or more people in the scene.

using the sensor fusion formulation alone or in combination with the reliability criterion. For the purpose of this experiment, we used the PILLARS sequence and manually segmented the occluder in each view for a ground truth comparison, and focused on a subregion of the scene in which the expected behaviors are well isolated. Fig. 6 shows that both schemes converge reasonably close to the visual hull of the considered pillar. In scenes with concave parts accessible to dynamic objects, the estimation would carve into concavities and reach a better estimate than the occluder's visual hull. A somewhat larger volume is reached with both schemes in this example. This is attributable to calibration errors which overtightens the visual hull with respect to the true silhouettes, and accumulation of errors in both schemes toward the end of the sequence. We trace those to the redundant, periodical poses contained in the video, that sustain consistent noise. This suggests the existence of an optimal finite number of frames to be used for processing. Jolts can be observed in both volumes corresponding to instants where the person walks behind the pillar, thereby adding positive contributions to the inference. Use of the reliability criterion contributes to lower sensitivity to noise, as well as a permanently conservative estimate of the occluder volume as the curves show in frames 100-200. Raw inference (10) momentarily yields large hypothetical occluder volumes when data is biased toward contributions of an abnormally low subset of views (frame 109).

5.4. Accounting for Occlusion in SfS

Our formulation (§4.2) can be used to account for the accumulated occluder information in dynamic shape inference. We only use occlusion cues from reliable voxels ($R > 0.8$) to minimize false positive occluder estimates, whose excessive presence would lead to sustained errors. While in many cases the original dynamic object formulation [6] performs robustly, a number of situations benefit from the additional occlusion knowledge (Fig. 7). Person volume estimates can be obtained when accounting for occluders. These estimates appear on average to be a stable multiple of the real volume of the person, which depends mainly on camera configuration. This suggests a possible biometrics application of the method, for disambiguation of person recognition based on computed volumes.

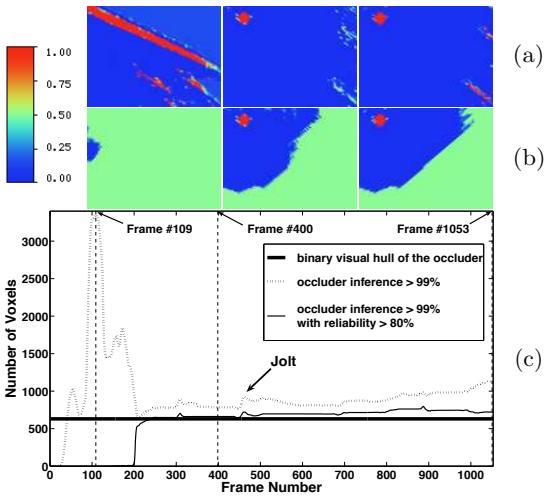

Figure 6. Online inference analysis and ground truth visual hull comparison, using PILLARS dataset, focusing on a slice including the middle pillar (**best viewed in color**). (a) Frames 109, 400 and 1053, inferred using (10). (b) Same frames, this time excluding zones with reliability under 0.8 (reverted here to 0.5). (c) Number of voxels compared to ground truth visual hull across time.

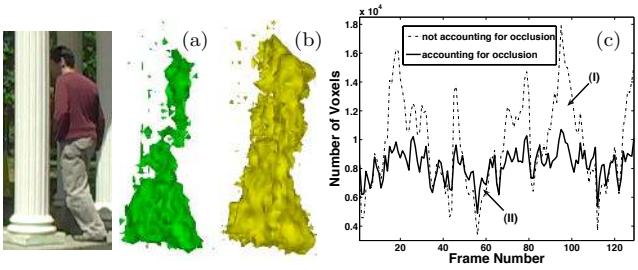

Figure 7. (a) Person shape estimate from PILLARS sequence, as occluded by the rightmost pillar and computed without accounting for occlusion. (b) Same situation accounting for occlusion, showing better completeness of the estimate. (c) Volume plot in both cases. Accounting for occlusion leads to more stable estimates across time, decreases false positives and overestimates due to shadows cast on occluders (I), increases estimation probabilities in case of occlusion (II).

6. Discussion

We have presented a method to detect and build the dense 3D shape of occluders indirectly observed through the motion of dynamic objects in a scene, in calibrated videos obtained from multiple views. The proposed Bayesian sensor formulation provides a useful probabilistic occluder representation, enabling detection and online accumulation of occluder information, and cooperative estimation of occluder and dynamic object shapes. The provided framework is robust to noise and avoids hard decisions about scene state. This new approach can lead to promising applications. Shape-from-occlusion could prove useful in conditions where segmenting objects is difficult or doesn't make sense, and using a moving object is easier, when all cameras don't have a complete view of the occluder, for example. Visual media such as infrared images exhibiting cold static objects of interest, inseparable from a broader cold background, could be used for modeling using a third, warm moving object. Detecting occlusions using this method can be helpful for a number of vision problems related to modeling, not limited to silhouette-based approaches. Many extensions are possible, such as automatic detection of changes in occluder configuration, cooperative background color model updates and occlusion estimation, integration of other cues such as color and texture.

References

[1] N. Apostoloff and A. Fitzgibbon. Learning spatiotemporal t-junctions for occlusion detection. *CVPR'05*, (2) p. 553–559.

[2] B. G. Baumgart. *Geometric Modeling for Computer Vision*. PhD thesis, CS Dept, Stanford U., Oct. 1974. AIM-249, STAN-CS-74-463.

[3] G. Cheung, T. Kanade, J.-Y. Bouguet, and M. Holler. A real time system for robust 3d voxel reconstruction of human motions. *CVPR'00*, vol. 2, p. 714–720.

[4] A. Elfes. Using occupancy grids for mobile robot perception and navigation. *IEEE Computer, Special Issue on Autonomous Intelligent Machines*, 22(6):46–57, June 1989.

[5] P. Favaro, A. Duci, Y. Ma, and S. Soatto. On exploiting occlusions in multiple-view geometry. *ICCV'03*, p. 479–486

[6] J.-S. Franco and E. Boyer. Fusion of multi-view silhouette cues using a space occupancy grid. *ICCV'05*, (2)1747–1753.

[7] B. Goldlücke and M. Magnor. Joint 3-d reconstruction and background separation in multiple views using graph cuts. *CVPR'03*, vol. 1, p. 683–694.

[8] B. Goldlücke and M. A. Magnor. Space-time isosurface evolution for temporally coherent 3d reconstruction. *CVPR'04*, vol. 1, p. 350–355

[9] K. Grauman, G. Shakhnarovich, and T. Darrell. A bayesian approach to image-based visual hull reconstruction. *CVPR'03*, (1)187–194.

[10] L. Guan, S. Sinha, J.-S. Franco, and M. Pollefeys. Visual hull construction in the presence of partial occlusion. *3DPVT'06*.

[11] A. Laurentini. The Visual Hull Concept for Silhouette-Based Image Understanding. *PAMI*, 16(2):150–162, Feb. 1994.

[12] D. Margaritis and S. Thrun. Learning to locate an object in 3d space from a sequence of camera images. *International Conference on Machine Learning*, p. 332–340, 1998.

[13] W. Matusik, C. Buehler, R. Raskar, S. Gortler, and L. McMillan. Image Based Visual Hulls. *Siggraph'00*, p. 369–374.

[14] D. Snow, P. Viola, and R. Zabih. Exact voxel occupancy with graph cuts. *CVPR'00*, p. 345–353.

[15] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. *CVPR'99*, (2) 246–252.

[16] R. Szeliski. Rapid Octree Construction from Image Sequences. *CVGIP*, 58(1):23–32, 1993.

[17] Z. Zhang. A flexible new technique for camera calibration. *PAMI*, 22(11):1330–1334, 2000.