Multi-Object Shape Estimation from Silhouette Cues
Submitted to ICCV 2007

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

This paper deals with the 3D shape estimation from silhouette cues of multiple moving objects in general 3D scenes, using video streams obtained from multiple calibrated cameras. Most silhouette-based methods usually assume that silhouette information corresponds to an underlying two-classification, based on image regions that match or disagree with a static background appearance model. However this reasoning becomes limited as the number of observed objects increases, because silhouette reasoning then becomes increasingly ambiguous and is unable to carve free space regions that happen to project inside real object silhouettes. We propose in this paper to distinguish between different objects in the scene by using object-specific appearance models. Reasoning can then be performed using a Bayesian formulation with multiple silhouette labels which decreases silhouette ambiguities. The estimation is reproduced iteratively to refine the models by accounting for occlusion relationships between different objects. Results show that shape estimates obtained using this method yield multiple silhouette-based estimates that drastically improve scene reconstructions over traditional two-label silhouette scene analysis.

1. Introduction

Silhouette-based methods have become increasingly popular in recent years for many computer vision problems such as 3D reconstruction, object localization and tracking, image-based modeling, shape analysis. This success is due to the simplicity, general robustness and speed of such methods to provide global shape and topology information about objects, which benefits a wide range of applications, such as 3D photography, human-machine interaction, markerless motion capture, etc. While silhouette-based shape modeling methods perform well in situations where the setup and light are controlled, and automatic background segmentation is easy, using such techniques can be difficult in more general conditions, where assumptions about visibility, lighting and scene content brake. Particularly multiple objects can make silhouette reasoning challenging and prone to large visual ambiguities, leading to misclassifications of significant portions of 3D space. Occlusion between dynamic objects of interest can be common and amplify the problem. Occlusions can also happen with static objects in the scene that make silhouette extraction difficult.

In this paper we show that silhouette reasoning can be efficiently conducted by using distinct appearance models for objects, yielding a multi-silhouette modeling approach. We propose a Bayesian framework to merge silhouette cues arising from a set of dynamic objects, and show that object occlusions and additional object localization constraints can be iteratively included in the modeling. This result is shown to improve shape-from-silhouette estimation, and can naturally benefit other vision problems such as multi-view tracking, segmentation, and general 3D modeling.

1.1. Previous work

Silhouette-based modeling in calibrated multi-view sequences has been largely popular, and yielded a large number of approaches to build volume-based [13] or surface-based [1] representations of the object’s visual hull. The difficulty and focus in attention in modeling objects from silhouettes has gradually shifted from the pure 3D reconstruction issue to the sensitivity of visual hull representations to silhouette noise. In fully automatic modeling systems, silhouettes are usually extracted using background subtraction techniques [12, 2], which are difficult to apply outdoors and often locally fail due to changing lighting conditions, shadows, color space ambiguities, background object induced occlusion, among other causes. Several solutions have been proposed to address these problems, using a discrete optimization scheme [11], silhouette priors over multi-view sets [6], or silhouette cue integration using a sensor fusion paradigm [4]. Most existing reconstruction methods however focus on mono-object situations, and fail to address the multi-object issues specific to silhouette methods.
While inclusive of the object’s shape [7], visual hulls fail to capture object concavities but are usually very good at hinting toward the overall topology of a single observed object, an empirical property that has been successfully used in a number of photometric-based methods to carve an initial silhouette-based volume [10, 5].

This ability to capture topologies breaks with the multiplicity of objects in the scene, where the sensitivity of silhouette-based methods to large non-convexities again becomes an issue. In such cases silhouettes alone can be ambiguous in distinguishing between regions actually occupied by objects and unfortunate silhouette-consistent “ghost” regions. Such regions have been analyzed in the context of tracking applications to avoid committing to a “ghost” track [9], for example. The method we propose casts the problem of silhouette modeling at the multi-object level, where ghosts can naturally be eliminated based on per object silhouette consistency.

Perhaps the closest related work is the approach of Ziegler et al. [15], which builds 3D models deterministically from multiple label, user-provided silhouette segmentations. The approach we propose only requires a prior model of each object’s appearance, and is able to produce probabilistic models automatically while accounting for process noise.

1.2. Principle

The ghost phenomenon occurs when the configuration of the scene is such that regions of space occupied by objects of interest cannot be disambiguated from free-space regions that also happen to project inside all silhouettes, as the polygonal gray region in Fig. 1(a). Ghosts are increasingly likely as the number of observed objects rises, because it then becomes more difficult to find views that visually separate objects in the scene and carve out unoccupied regions of space. The problem is not entirely specific to silhouettes because it relates to the more general occlusion and visibility problem, which other types of shape modeling techniques must account for.

To address this problem, we propose to enrich the silhouette representation based on multi-object silhouette cues, by learning a set of appearance models associated to m objects in the scene, either from pre-observed sequences or online, as well as the pixel-appearance of static backgrounds. The intuition is then that the probability of confusing ambiguous regions with real objects decreases, because the silhouette set corresponding to ghosts is then drawn from non object-consistent appearance model sets, as depicted in Fig. 1(b).

It is possible to process multiple silhouette labels in a deterministic, purely geometric fashion [15], but this comes to the price of an arbitrary hard decision through a threshold to the number of views that define consistency. Silhouettes are then also assumed to be manually given and noiseless, which cannot be assumed for automatic processing. We thus process multi-object sequences by examining each voxel in the scene using a Bayesian formulation, which encodes the noisy causal relationship between the voxel and the pixels that observe it in a generative sensor model. In particular, given the knowledge that a voxel is occupied by a certain object among m possible in the scene, the sensor model explains what appearance distributions we are supposed to observe, corresponding to that object. It also encodes state information about the viewing line and potential obstructions from other objects, as well as a localization prior used to enforce the compactness of objects, which can be used iteratively to refine the estimate for a given instant of the sequence. Voxel sensor model semantics and simplifications are borrowed to the occupancy grid framework explored in the robotics community [3, 8]. The proposed method can also be seen as a multi-object generalization of previous probabilistic approaches focused on 2-label silhouette modeling [4].

This scheme enables us to perform silhouette inference in a way that reinforces regions of space which are drawn from the same conjunction of color distributions, corresponding to one object, and penalizes appearance inconsistent regions, while accounting for object visibility. It also imposes no hard decision because the output representation is a per-voxel probability distribution over the labels which naturally encode the notion of noisy object consistency.

2. Formulation

We consider a scene observed by n calibrated cameras. We assume a maximum of m dynamic objects of interest can be present in the scene. In this formulation we focus on the state of one voxel at position X chosen among the positions of the 3D lattice used to discretize the scene. We here model how knowledge about the occupancy state of voxel X influences image formation, assuming a static appearance model for the background has previously been observed. Because of occlusion relationships arising between objects, the zones of interest to infer the state of voxel X are its n viewing lines \( L_i, i \in \{1, \ldots, n\} \), with respect to the different views. In this paragraph we assume that some prior knowledge about scene state is available for each voxel
X in the lattice and can be used in the inference. Various uses of this assumption will be demonstrated in §3 and §4. A number of statistical variables are used to model the state of the scene, depicted in figure Fig. 2, are used to model the image generation process and infer \( \mathcal{G} \).

### 2.1. Statistical Variables

**Scene voxel state space.** The occupancy state of \( X \) is represented by a variable \( \mathcal{G} \). The particularity of our modeling lies in the multi-labeling characteristic of \( \mathcal{G} \in \mathcal{L} \), where \( \mathcal{L} \) is a set of labels \( \{ \emptyset, 1, \cdots, m, \mathcal{U} \} \). A voxel is either empty (\( \emptyset \)), or one of \( m \) objects the model is keeping track of (numerical labels), or occupied by an unidentified object (\( \mathcal{U} \)). \( \mathcal{U} \) is intended to act as a default label capturing all objects that are detected as different than background but not explicitly modeled by other labels, which proves useful for automatic initialization of appearance models (§4).

**Observed appearance.** The voxel \( X \) projects to a set of pixels, whose colors \( \mathcal{I}_i \), \( i \in 1, \cdots, n \) we observe in images. We assume these colors are drawn from a set of object and view specific color models whose parameters we note \( C_i^l \). More complex appearance models are possible using texture information, or color models that account for spatial correlations and neighborhoods in the image. We here use a simple model to demonstrate the key ideas of the framework and show its potential.

**Latent viewing line variables.** To account for inter-object occlusion, we need to model the contents of viewing lines and how it contributes to image formation. We assume some a priori knowledge about where objects lie in the scene. The presence of such objects can have an impact on the inference of \( \mathcal{G} \) because of the visibility of objects and how they affect \( \mathcal{G} \). Intuitively, conclusive information about \( \mathcal{G} \) cannot be obtained from a view \( i \) if a voxel in front of \( \mathcal{G} \) with respect to \( i \) is occupied by another object, for example. However, \( \mathcal{G} \) directly influences the color observed if it is unoccluded and occupied by one of the objects. But if \( \mathcal{G} \) is known to be empty, then the color observed at pixel \( \mathcal{I}_i \) reflects the appearance of objects behind \( X \) in image \( i \), if any. These visibility intuitions need to be accounted for in our modeling.

It is not meaningful to account for the combinatorial number of occupancy possibilities along the viewing rays of \( X \). This is because neighboring voxel occupancies on the viewing line usually reflect the presence of a same object and are therefore correlated. The fundamental information that is required to reason about \( X \) is the knowledge of presence and ordering of the \( m \) objects along this viewing line. To represent this knowledge, assuming prior information about occupancies is already available at each voxel, we extract, for each label \( l \in \mathcal{L} \) and each viewing line \( i \in \{1, \cdots, n\} \), the voxel whose probability of occupancy is dominant for that label on the viewing line. This corresponds to electing the voxels which best represent the \( m \) objects and have the most influence on the inference of \( \mathcal{G} \). We then account for this knowledge in the problem of inferring \( X \), by introducing a set of statistical occupancy variables \( \mathcal{G}_i^l \in \mathcal{L} \), corresponding to these extracted voxels.

### 2.2. Dependencies Considered

We propose a set of simplifications in the joint probability distribution of the set of variables, that reflect the prior knowledge we have about the problem. To simplify the writing we will often note the conjunction of a set of variables as following:

\[
\mathcal{G}_i^{1:m} = \{ \mathcal{G}_i^l \}_{l \in \{1, \cdots, m\}, i \in \{1, \cdots, n\}}.
\]

We propose the following decomposition for the joint probability distribution \( p(\mathcal{G}, \mathcal{G}_i^{1:m}, \mathcal{I}_1:n, C_i^{1:m}) \):

\[
p(\mathcal{G}) \prod_{l \in \mathcal{L}} p(C_i^{l;1:n}) \prod_{i, l \in \mathcal{L}} p(\mathcal{G}_i^l | \mathcal{G}) \prod_i p(\mathcal{I}_i | \mathcal{G}, \mathcal{G}_i^{1:m}, C_i^{1:m}) \quad (1)
\]

**Prior terms.** \( p(\mathcal{G}) \) carries prior information about the current voxel. This prior can reflect different types of knowledge and constraints already acquired about \( \mathcal{G} \), such as a localization or tracking prior, as will be further discussed in §3 and §4.

\( p(C_i^{l;1:n}) \) is the prior over the view-specific appearance models of a given object \( l \). The prior, as written over the conjunction of these parameters, could express expected relationships between the appearance models of different views, which could prove useful in the context of non-color
calibrated cameras. Since the focus is on the learning of voxel $X$, we do not use this capability here and assume $G_{1:n}$ to be uniform.

**Viewing line dependency terms.** We have summarized the prior information along each viewing line using the $m$ voxels most representative of the $m$ objects, so as to model inter-object occlusion phenomena. However when examining a particular label $G = l$, keeping the occupancy information about $G_l$ would lead us to account for intra-object occlusion phenomena, which in effect would lead the inference to favor mostly voxels from the front visible surface of the object $l$. Because we wish to model the volume of object $l$, we discard the influence of $G_l$ when $G = l$:

$$p(G^k_l|G = l) = P(G^k_l) \quad \text{when } k \neq l \quad (2)$$

$$p(G^k_l|G = l) = \delta_l(G^k_l) \quad \forall l \in L, \quad (3)$$

where $P(G^k_l)$ is a distribution reflecting the prior knowledge about $G^k_l$, and $\delta_l(G^k_l)$ is the distribution giving all the weight to label $\emptyset$. In (3) $p(G^k_l|G = l)$ is thus enforced to be empty when $G$ is known to be representing label $l$, which ensures that $G^k_l$ has no influence on the inference in this case.

**Image formation terms.** The image formation term $p(I_i|G, G_i^{1:m}, C_i^{1:m})$ explains what color we expect to observe given the knowledge of viewing line states and per-object color models. We decompose each such term in two subterms, by introducing a local latent variable $S \in L$ representing the hidden silhouette state:

$$p(I_i|G, G_i^{1:m}, C_i^{1:m}) = \sum_{S} p(I_i|S, C_i^{1:m}) p(S|G, G_i^{1:m}) \quad (4)$$

The term $p(I_i|S, C_i^{1:m})$ simply describes what color is likely to be observed in the image given the knowledge of the silhouette state and the appearance models corresponding to each object. $S$ acts as a mixture label: if $\{S = l\}$ then $I_i$ is drawn from the color model $C_i^l$. For objects ($l \in \{1, \ldots, m\}$) we typically use GMMs [12] to efficiently summarize the appearance information of dynamic object silhouettes. For background ($l = \emptyset$) we use per-pixel Gaussians as learned from pre-observed sequences, although other models are possible. When $l = \emptyset$ the color model is drawn from the uniform color distribution, as we can make no assumption about the color of previously unobserved objects.

Defining the silhouette formation term $p(S|G, G_i^{1:m})$ requires that the variables be considered in their visibility order on the viewing lines, in order to model the occlusion possibilities. Note that this order can be different from $1, \cdots, m$. We note \{\$G^k_{i_{j}}\}_{j \in \{1, \cdots, m\}}$ the variables $G^{1:m}$ as enumerated in the permuted order \{\$v_j\}_{i}$, reflecting their visibility ordering on $L_i$. If \{\$v_j\}_{i}$ denotes the particular index after which the voxel $X$ itself appears on $L_i$, then we can re-write the silhouette formation term as $p(S|G^1_{i_{1}}, \cdots, G^{V_{y_{s}}}_{i_{s}}, G^{V_{y_{s+1}}}_{i_{s+1}}, \cdots, G^{V_{y_{m}}}_{i_{m}})$. It then becomes easy to assign a distribution to this term with the following generic form:

$$p(S|\emptyset \cdots \emptyset l \cdots \emptyset) = d_l(S) \quad \text{with } l \neq \emptyset \quad (5)$$

$$p(S|\emptyset \cdots \emptyset) = d_{\emptyset}(S), \quad (6)$$

where $d_k(S), \ k \in L$ is a family of distributions giving strong weight to label $k$ and lower equal weight to others, determined by a constant probability of detection $P_d \in [0, 1]$: $d_k(S = k) = P_d$ and $d_k(S \neq k) = \frac{1-P_d}{|L|-1}$ to ensure summation to 1. (5) thus expresses that the silhouette pixel state reflects the state of the first visible non-empty voxel on the viewing line, regardless of the state of voxels behind it (\$\ast\$). (6) expresses the particular case where no occupied voxel lies on the viewing line, which is the only case for which the state of $S$ should be background, thus ensuring that $I_i$ is mostly drawn from the background appearance model.

**2.3. Inference**

Estimating the occupancy at voxel $X$ translates to estimating $p(G|I_{1:n}, C_i^{1:m})$ in Bayesian terms. We apply Bayes’ rule using the joint probability distribution, marginalizing out the unobserved variables $G_i^{1:m}$:

$$p(G|I_{1:n}, C_i^{1:m}) = \frac{1}{z} \sum_{G_i^{1:m}} p(G, G_i^{1:m}, I_{1:n}, C_i^{1:m})$$

$$= \frac{1}{z} \prod_{i=1}^{n} f_i^l$$

$$= \sum_{G_i^{1:m}} p(I_i|G, G_i^{1:m}) p(S|G, G_i^{1:m}) \quad (7)$$

$$\text{where } f_i^l = \sum_{G_i^{1:m}} p(I_i|G, G_i^{1:m}) \quad \text{for } k < m \quad (9)$$

$$\text{and } f_i^m = \sum_{G_i^{1:m}} p(I_i|G, G_i^{1:m}) \quad \text{for } k = m \quad (10)$$

The normalization constant $z$ is easily obtained by ensuring summation to 1 of the distribution: $z = \sum_{G_i^{1:m}} p(G, G_i^{1:m}, I_{1:n}, C_i^{1:m})$. (7) is the direct application of Bayes rule, with the marginalization of unknown variables. This sum in this form is intractable, thus we factorize the sum in (8). The sequence of $m$ functions $f_i^l$ specify how to recursively compute the marginalization with the sums of individual $G_i^l$ variables appropriately subsumed, so as to factor out terms not required at each level of the sum. Because of the particular form of silhouette terms in (5), this sum can be efficiently computed
by noting that all terms after a first occupied voxel of the
same visibility rank $k$ share a term of identical value in
$p(I_i \mid \emptyset \cdots \emptyset \{G_i^v = l\} \cdots \emptyset) = P_i(I_i)$. They can be
factored out of the remaining sum, which sums to 1 being
a sum of terms of a probability distribution, leading to the
following simplification of (9), $\forall k \in \{1, \cdots , m - 1\}$:

$$f_i^k = p(G_i^v = \emptyset | G) f_i^{k+1} + \sum_{l \neq \emptyset} p(G_i^v = l | G) P_i(I_i) \tag{11}$$

3. Refinement

In the previous section we have presented a generic
framework to infer the occupancy probability of a voxel $X$
and thus deduce how likely it is for $X$ to belong to one of
$m$ objects, assuming some prior knowledge about the scene
occupancies and viewing line states. The availability of this
prior information however implicitly assumes that part of
the occupancy problem was solved or that good hints were
acquired. Since we make no a priori assumptions about ob-
ject localization, this information can only be given by the
inference itself. Fortunately we can break the chicken and
egg problem, because bare multi-label silhouette inference,
striped of any processing of occlusion cues or space prior,
is already very robust in indicating the region of occupancy
of the $m$ objects based on appearance information alone, as
confirmed in §5. Thus this bare inference can be used to ini-
tialize our modeling. It is a particular case of the problem,
where no knowledge is assumed about viewing line obstruc-
tions. As a simplification, each voxel can be treated as di-
rectly observed and unobstructed, leading to the following
trivial reduction of the inference (8):

$$p(G | I_{1:n} C_{1:m}^i) = \frac{1}{2} p(G) \prod_{i=1}^n p(I_i | G C_{1:m}^i) \tag{12}$$

Nevertheless, this first solution is coarse when there are
a high number of objects in the scene. We thus opt for
an iterative refinement approach, where the results of the
previous pass are used to initialize the voxel priors $p(G)$
and $P(G_j)$ of equation (2) for the next pass. This enables
the method to deal with occlusion information after the
initialization pass.

Object Compactness Prior. We use a localization prior
to enforce the compactness of objects from one iteration
to another. For the particular sequences we processed (walk-
ing people), we opt for a solution that takes advantage of the
underlying structure of the dataset, by projecting the maxi-
mum probability over a voxel column on the horizontal ref-
ence plane. We then localize the most likely position of
objects by fixed-size sliding a window over the resulting 2D
probability map for each object. The resulting center is then
used to initialize $p(G)$, using a Gaussian spatial prior. This
heuristic thus favors objects localized in one and only one
portion of the scene and is intended as a soft guide to the
inference.

However we do not use this method to initialize $P(G_j)$,
because we wish to capture finer detail in the 3D occlusions:
we thus initialize $P(G_j)$ to be exactly the result of the pre-
vious inference. Naturally other specific choices are possible
and could be used in different modeling contexts.

4. Automatic Initialization

The main information about objects used by the pro-
posed method is their set of appearances in the different
views. These sets can be learned offline by segmenting each
observed object alone in a clear, uncluttered scene before
processing multi-objects scenes. This is however a non-
generic solution to the problem. Another possibility we
have explored is to provide automatic initialization of ob-
ject color models in the scene. To reduce the difficulty of
the problem, we have focused on the case where new ob-
jects arrive in the scene one at a time from an unoccluded
entry point to the scene when objects are already present. To
be able to learn appearance models of new objects in this
context, one needs to recognize that a new object, which
doesn’t correspond to anything already labeled in the scene,
has appeared in the sequence. We use an additional label $U$
partly of the label set $\mathcal{L}$, which captures most non-empty
voxels that are not identified by the existing objects color
models. We project the maximum probabilities vertically
to determine the likeliest region that is unmodeled by slid-
ing a window over the map. If above a certain detection
threshold, the combined probabilities of the best candidate
window determines whether the unmodeled region is con-
 sidered a new object. Then the voxels of the unmodeled
region are reprojected on the source image and the appear-
ance models are learned using the provided pixel examples.
We show results with this scheme in §5.

5. Results

We have performed two multi-view acquisitions for the
purpose of validation: the CLUSTER and WALKING se-
cence. They are acquired outdoors with eight 30fps DV
cameras surrounding the scene in a semi-circle, with im-
age resolutions of $720 \times 480$. Both sequences are par-
cularly difficult from a reconstruction and localization per-
spective, because both are densely populated. The CLUS-
TER sequence provides a particularly hard configuration in
that the five people are placed on a circle of less than 3m
in diameter, yielding an extremely ambiguous and occluded
situation in the center of the circle. Despite the fact that
none of them are being observed in all views, we are still
able to recover the people’s label and shape. We provide a supplemental video with additional results.

Cameras in each data sequence are geometrically calibrated, using Bouguet’s toolbox based on [12], and recording at 30Hz. Color calibration is unnecessary because the model uses per-view 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, window reflections and substantial illumination changes, illustrating the robustness of the method to difficult real conditions.

For appearance models of dynamic objects of the CLUSTER sequence, we train an RGB GMM model for each person in each view (because of non-color calibration), using a manually segmented image from that view. We typically use a dozen modes for the GMM.

Processing was handled on a 2.8 GHz PC. For each time step, we choose to perform a fixed number of passes. We achieve approximately 4 minutes per time step with 3 passes, for both the CLUSTER and WALKING datasets (lattice size $128^3$). The very strong locality inherent to the algorithm and preliminary benchmarks suggest that around 10 times faster performance could be achieved using a GPU implementation.

5.1. Densely populated scene

We here analyze the results of the CLUSTER sequence, with results depicted in Fig. 3. The figure shows the improvement given by each different component in our model, and provides a comparison with traditional 2-label silhouette analysis. The 2-label reconstruction yields one large volume with no separation between objects, because the entire scene configuration is too ambiguous (Fig. 3-1). The naive, one pass, multi-label approach doesn’t use any localization prior or occlusion cue inference (Fig. 3-2), but still succeeds in giving a rough estimate for each label. However large artifacts are to be seen: the heads are misclassified because the yellow color model comes from a person with a black T-shirt which attracts all black colors such as hair. Adding prior knowledge about localization and compactness helps to tighten the estimate around the highest probability region and eliminate large errors, to the expense of dilatation and lack of precision (Fig. 3-3). Accounting for viewing line occlusions enables the model to recover large chunks of information, including some limbs, showing a large improvement over all other modeling attempts (Fig. 3-4).

5.2. Automatic appearance model initialization

An automatic approach for silhouette appearance model initialization has also been tested using the WALKING sequence, as shown in Fig. 4. Three people are initially accounted for in the modeling, and a fourth person (with orange shirt) walks into the scene, being momentarily classified as “unmodeled”. As soon as its probability becomes significant the model is detected and is instantiated as a new object, with its appearance model learned.

6. Discussion

We have proposed a Bayesian method to build 3D shapes from multi-object silhouette cues. The appearances of objects are used to disambiguate free regions of space which project inside silhouettes, and occlusion information and object localization priors are used to update the representation iteratively so as to refine the resulting shapes. Our results show that the shapes obtained using this approach yield significantly better results than pure silhouette reasoning, which makes no distinction between different objects. This new multi-silhouette inference algorithm is robust to very difficult conditions, and can prove very useful for various vision tasks such as tracking, localization and 3D reconstruction, in highly cluttered scenes with densely packed dynamic object groups. A large number of extensions can be tested on the basis of the framework provided, including more general and complex appearance modeling, different enforcements of the compactness of objects, a more general management of objects entering and leaving the scene. It is possible to include static occluders (such as the bench) in the framework as it is a different type of object, whose appearance is recorded as background and could model additional occlusion phenomena. Temporal consistency constraints can also be included in various forms: terms can be added to enforce temporal continuity of the reconstruction and smoothness of the flow in the scene.

References


Figure 3. Probabilistic object occupancy labeling of five clustered people in an outdoor environment. Three of the total eight camera views are shown on the right column. On the top left, horizontal slices of the same inferred voxel grids are shown, approximately at waist level. (1) inference with 2-label reasoning, (2) multi-silhouette inference with no localization prior or occlusion cues, (3) multi-silhouette inference using a localization prior. (4) adds an additional pass to the result of (3) accounting for occlusion cues. The reconstructed result thresholded at probability 0.8 are shown in the bottom, with one label color per person. See supplemental video for extensive results. Best viewed in color.

Figure 4. Two views among the 8 used are shown on the right column. (1) shows three people currently being modeled and a gray region corresponding to a fourth person about to be detected. Only raw inference and localization priors are used at this stage (2) shows the next inference pass, which accounts for the newly created appearance model set, and adds occlusion reasoning. The new person is correctly reconstructed and the models stabilized and refined thanks to occlusion cues.


