Fusion of Multi-View Silhouette Cues Using a Space Occupancy Grid

Jean-Sébastien Franco Edmond Boyer
GRAVIR - INRIA Rhône-Alpes
655, av. de l'Europe, 38330 Montbonnot, France
firstname.lastname@inrialpes.fr

Abstract

In this paper, we investigate what can be inferred from several silhouette probability maps, in multi-camera environments. To this aim, we propose a new framework for multi-view silhouette cue fusion. This framework uses a space occupancy grid as a probabilistic 3D representation of scene contents. Such a representation is of great interest for various computer vision applications in perception, or localization for instance. Our main contribution is to introduce the occupancy grid concept, popular in the robotics community, for multi-camera environments. The idea is to consider each camera pixel as a statistical occupancy sensor. All pixel observations are then used jointly to infer where, and how likely, matter is present in the scene. As our results illustrate, this simple model has various advantages. Most sources of uncertainty are explicitly modeled, and no premature decisions about pixel labeling occur, thus preserving pixel knowledge. Consequently, optimal scene object localization, and robust volume reconstruction, can be achieved, with no constraint on camera placement and object visibility. In addition, this representation allows to improve silhouette extraction in images.

1. Introduction

Silhouette-based methods are popular for use in multi-camera environments mainly due to their simplicity and computational efficiency. These methods concern 3D modeling, multi-object localization and motion capture applications, among others. Often however in such methods, silhouettes of objects of interest are extracted using a binary labeling of pixels into foreground or background, for each view separately, and prior to any 3D operation. Unfortunately, such monocular labeling, called *background subtraction*, is difficult to achieve in a general and uncontrolled environment. Several reasons account for this, in particular perturbations due to: camera sensor noise, ambiguities between objects and background colors, changes in the lighting of the scene (including shadows of objects of interest), etc. In addition, monocular background labeling can dramatically alter 3D perception from multiple views in the presence of camera calibration errors, or if disparities between image acquisition times exist.

Our goal is therefore to find a representation of multi-view silhouette cues, where inference about silhouettes is of greater robustness to the aforementioned uncertainties than single view silhouette inference. Intuitively, the simultaneous knowledge of all images brings more information about silhouettes than knowledge from only one image. This idea has lead us to compute silhouette fusion in 3D space, in order to integrate the contribution of all images. The result of such fusion naturally encodes shape information. As such it can be used to improve many silhouette-based applications, from shape modeling to silhouette extraction, as we will show.

Very often silhouettes are used to infer shapes in a two-step process: an individual decision about silhouette occupancy is made on a per-view basis, then shape and position are inferred geometrically from all available silhouettes using *visual hull* methods [11]. These methods can lead to a surface representation of the objects of interest [5], a voxel representation [16], or image-based representation [13]. While visual hull estimation can be exact from a set of silhouettes [5], silhouette extraction methods come generally with several caveats resulting from the perturbations mentioned earlier. Our approach allows to delay the occupancy decision to a later stage and, as such, makes a better use of the available silhouette information.

Several methods have also been proposed to bypass silhouette estimation altogether, as many algorithms reconstruct the scene structure based only on photometric information [10]. Others possibly state it as the solution of a global state optimization problem: using level sets [4], or graph cuts [7]. Probability grid representations have already been used by the community, mainly to solve photometric problems[1]. These methods generally have high complexities and computational costs compared to silhouette methods, as they must deal with the visibility of points on the object's surface. This is why there are still many situations

where silhouette methods are preferred (e.g. VR platforms, real-time setups), or used to initialize a more elaborate photometric method [9].

More closely related, Magnor *et al.* [7] solve a similar problem with two views using graph-cuts, where stereo disparity and silhouettes are simultaneously estimated. Zeng *et al.* propose a multi-view background silhouette extraction, based on a costly geometric scheme, with the additional constraint of common object visibility [18]. A similar idea for silhouette information integration has been proposed, using however a discrete formulation and a coarser image model [14]. Grauman *et al.* [8] propose a method to estimate the most probable multi-view silhouette set using a learned human silhouette prior, and therefore integrate a higher level of semantics, but with limited genericity. Robotics works from S. Thraun *et al.* [12] propose a solution for the closely related problem of object localization from a robot-acquired image sequence. These approaches solve silhouette-based problems in multi-view sequences with, however, limited application domains. Our approach is at a lower level, and is intended to enrich 2D silhouettes cues by embedding them into a 3D representation independently of the application.

We propose a new framework based on the occupancy grid: a voxel grid of object occupancy probabilities in space, associated to a sensor model. The occupancy grid has been extensively used in the robotics community [3], to represent a robot's environment for navigation, based on range sensor observations, with depth and orientation measurements. Our contribution is to extend the occupancy grid concept to image sensors, and to restate shape-from-silhouette estimation as a sensor fusion problem. To this extent, we provide each pixel with a *forward* sensor formulation which models the pixel observation responses to the voxel occupancies in the scene. Our formulation accounts for each pixel's visibility region, voxel sampling issues, small camera calibration errors, and sensor reliability. This model is in turn used to infer the answer to the more difficult inverse question: given the color observations, where is the matter located in the scene. We also show that the resulting occupancy grid can be used to perform multi-view background subtraction, where silhouette estimation in each view benefits from the knowledge of other views.

2. Problem Statement

We consider the problem of silhouette cue fusion from multiple views. We assume we are given a *current set* of images, obtained from fully calibrated cameras. We also assume that a set of *background images* of the scene, free from any *object of interest*, have previously been observed for each camera. Importantly, no assumption is made about the existence of a visibility domain common to all cameras.

The problem is formulated as the separate Bayesian estimation, for each voxel, of how likely it is occupied by an object of interest. We formulate the problem using a forward sensor model: we model the relationship from causes to observations. Namely, in our problem, we will model how a voxel influences image formation. This enables us, using Bayesian inference, to solve the more difficult inverse problem: express the voxel occupancy likelihood using images as a noisy measurement of scene state.

Solving a Bayesian problem requires computing the joint probability of all variables of interest (which we define in §2.1), prior to any inference. This joint probability distribution must then be decomposed and simplified, based on the main statistical dependencies we choose to consider between variables (§3 and §3.1). In particular, parametric forms must be assigned to the various terms of the decomposition to explicitly model the uncertain relationship between variables (§3.2 and §3.3). This simplifies the inference of voxel occupancy distributions, which are inferred from the joint probability expression using Bayes' rule (§4).

2.1. Main problem variables

We label the set of n current images as $\mathcal{I}$. $\mathcal{I}^i$, $i = 1 \cdots n$ is then the image data of camera i, and $\mathcal{I}^i_p$ is the image data at pixel p in image i, expressed in some color space (RGB, YUV, etc). Although not studied explicitly in this paper, additional image cues can be enclosed in the $\mathcal{I}^i_p$ term, such as the image gradient or some other local feature, without loss of generality. We assume that the image data of the corresponding m observed background images can be summarized into a single statistical model image $\mathcal{B}^i$, $i = 1 \cdots n$. Both image data sets are produced by n cameras with known projection matrices $\mathbf{P}^i$. τ symbolizes the prior knowledge we introduce into the model. This includes what we now about the scene, what we know about sensor characteristics, our general knowledge about the system.

We define $\mathcal{G}$ as our space occupancy grid. For each space point X in the grid discretization we associate the corresponding binary occupancy variable $\mathcal{G}_X \in \{0, 1\}$, respectively free or occupied. As a common occupancy grid assumption [3], we assume statistical independence between voxel occupancies, and compute each voxel occupancy likelihood independently for tractability. Results show that independent estimation, while not as exhaustive as a global search over all voxel configurations, still provides very robust and usable information, at a much lower cost.

We have defined our input and output variables. We now introduce an important hidden variable set per image, the silhouette detection maps $\mathcal{F}^i$, $i = 1 \cdots n$. These maps define, for each pixel p in image i, a binary silhouette detection variable $\mathcal{F}^i_p$. $\mathcal{F}^i_p = 1$ if the pixel sensor p in image i *reports* the presence of an object of interest anywhere along its viewing line. We insist on this definition, since there is a

possibility that an object *is* indeed present along the viewing line of pixel p, but that the pixel sensor itself *fails* to detect and report this information for internal or external causes (modeling sensor failures will be discussed in §3.2). These detection maps represent the silhouette information in our model, over which we wish to marginalize.

3. Joint Probability Decomposition

Our goal is to infer the occupancy $\mathcal{G}_X$ of a voxel at position X, given $\mathcal{I}$, $\mathcal{B}$, and τ. Thus, we must first model the impact of $\mathcal{G}_X$ on the observations. Modeling the relationships between the variables involved requires computing the joint probability of these variables, $p(\mathcal{G}_X, \mathcal{I}, \mathcal{B}, \mathcal{F}, \tau)$. We propose the following decomposition, based on the statistical dependencies expressed in Fig. 1:

$$p(\mathcal{G}_X, \mathcal{I}, \mathcal{B}, \mathcal{F}, \tau) = p(\tau)\ p(\mathcal{B}\,|\,\tau)\ p(\mathcal{G}_X\,|\,\tau)\ p(\mathcal{F}\,|\,\mathcal{G}_X, \tau)\ p(\mathcal{I}\,|\,\mathcal{F}, \mathcal{B}, \tau)$$

- $p(\tau)$, $p(\mathcal{B}\,|\,\tau)$ are the prior probabilities of our parameter set, and of background image parameters. Since we have no *a priori* reason to favor any parameter values, or background image configurations, we set these terms to a uniform distribution. They thus disappear from any subsequent inference.
- $p(\mathcal{G}_X\,|\,\tau)$ is the prior likelihood for occupancy, which could vary with X for example. We consider the occupancy to be at the top of the causality chain, thus the independence with all other variables except τ. We choose not to favor any voxel location and set this term to uniform, being mainly interested in the regularization of voxels induced by observations in this paper.
- $p(\mathcal{F}\ |\ \mathcal{G}_X, \tau)$ is the silhouette likelihood term. The dependencies considered reflect that voxel occupancy in the scene explains object detection in images.
- $p(\mathcal{I}\,|\,\mathcal{F}, \mathcal{B}, \tau)$ is the image likelihood term. Image colors are conditioned by object detections in images, and the knowledge of the background color model.

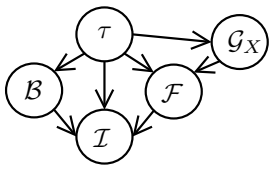

Figure 1. Variables of our system and their dependency graph. τ: prior knowledge we introduce in the model. $\mathcal{G}_X$: occupancy at voxel X. $\mathcal{B}$: background model maps. $\mathcal{F}$: silhouette detection maps. $\mathcal{I}$: observed images.

3.1. Sensor fusion simplifications

Pixel colors in input images are treated as noisy observations of the model. We consider that the noise is independently and identically distributed. Each pixel's color observation can be considered independent of all others, given the observation's main cause, the background data and silhouette detection state of the pixel: the image likelihood term can thus be simplified to a product of per pixel terms, $p(\mathcal{I}\,|\,\mathcal{F}, \mathcal{B}, \tau) = \prod_{i,p} p(\mathcal{I}_p^i\,|\,\mathcal{F}_p^i, \mathcal{B}_p^i, \tau)$.

All pixel detections can also be considered independent, given the knowledge of their main cause, namely the voxel occupancy. The silhouette likelihood is therefore similarly simplified: $p(\mathcal{F}\ |\ \mathcal{G}_X, \tau) = \prod_{i,p} p(\mathcal{F}_p^i\ |\ \mathcal{G}_X, \tau)$. Thus, the joint probability distribution of variables of interest reduces to the following product of per pixel terms:

$$p(\mathcal{G}_X, \mathcal{I}, \mathcal{B}, \mathcal{F}, \tau) = \prod_{i,p} p(\mathcal{F}_p^i\,|\,\mathcal{G}_X, \tau) p(\mathcal{I}_p^i\,|\,\mathcal{F}_p^i, \mathcal{B}_p^i, \tau) \quad (1)$$

We have therefore reduced the evaluation of the joint probability of all variables to two much friendlier subproblems. First, expressing the likelihood of silhouette detection at a single pixel, given the knowledge of our voxel's occupancy (§3.2). Second, expressing the likelihood of the color observation at a single pixel, given the silhouette detection state, and background color information at this pixel (§3.3). We now focus on these two terms.

3.2. Silhouette Formation Term

The silhouette detection likelihood $p(\mathcal{F}_p^i\ |\ \mathcal{G}_X, \tau)$ models the silhouette detection response of a single pixel sensor (i, p) to the occupancy state of our voxel of interest $\mathcal{G}_X$. We need to introduce two local hidden process variables $\mathcal{S}$ and $\mathcal{R}$ to balance the influence of this voxel. Fig. 2 introduces the variables and statistical dependencies of this subproblem. In an ideal and noiseless setup, the two variables $\mathcal{F}_p^i$ and $\mathcal{G}_X$ would be self-sufficient and the relationship between them expressed as simple logic: if our voxel X is occupied, and if it projects to pixel p, then silhouette detection occurs at pixel p, $\mathcal{F}_p^i = 1$. This is the implicit formulation used by all classical visual hull methods.

However, there are sources of uncertainty which perturb this intuitive reasoning. First, the assumption that a voxel lies on the viewing line of a pixel is itself uncertain. This can be due to many external causes: potential camera calibration errors, camera mis-synchronization, which both introduce misalignment in the scene. Voxel sampling is also an issue, since no voxel perfectly projects to a pixel, and its projected surface can cover several. Second, there can be causes for sensor detection other than the voxel itself: an object occupancy other than the one related by $\mathcal{G}_X$, or a change in background scene appearance (an *internal* sensor failure due to the nature of the sensor model).

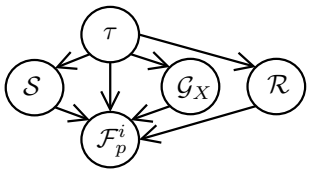

Figure 2. Variables and dependency graph of the per-pixel silhouette detection subproblem. τ: prior knowledge. $\mathcal{G}_X$: voxel occupancy. $\mathcal{S}$: sampling variable. $\mathcal{R}$: external detection cause. $\mathcal{F}_p^i$: silhouette detection at pixel (i,p).

Modeling these hidden causes is possible using two boolean random variables, the *sampling variable* $\mathcal{S}$ and *external detection cause variable* $\mathcal{R}$. This leads to two expressions for the silhouette detection prior $p(\mathcal{F}_p^i \,|\, \mathcal{G}_X, \tau)$. First, let us consider the case where our voxel X is known to be occupied ($\mathcal{G}_X = 1$):

$$\begin{aligned} p(\mathcal{F}_p^i \,|\, [\mathcal{G}_X{=}1], \tau) \;&=\; p(\mathcal{S}{=}0 \,|\, \tau)\; \mathcal{U}(\mathcal{F}_p^i) \\ &+\; p(\mathcal{S}{=}1 \,|\, \tau)\; \mathcal{P}_d(\mathcal{F}_p^i) \end{aligned} \tag{2}$$

By definition, $\mathcal{S}{=}1$ if voxel X is on the viewing line of pixel (i,p). When this is not the case ($\mathcal{S}{=}0$), the knowledge of our voxel's occupancy is irrelevant to sensor detections at this pixel, thus the uniform distribution $\mathcal{U}(\mathcal{F}_p^i)$ for silhouette detection in (2). If the voxel is on the viewing line of p ($\mathcal{S}{=}1$), then detection at the pixel is ruled by the probability distribution $\mathcal{P}_d(\mathcal{F}_p^i)$. In practice we set this distribution using a constant $P_D \in [0,1]$, which is a parameter of our system: $\mathcal{P}_d([\mathcal{F}_p^i = 1]) = P_D$ is the detection rate of a pixel sensor, and $\mathcal{P}_d([\mathcal{F}_p^i = 0]) = 1 - P_D$ is its detection failure rate. Detection failure occurs when the pixel sensor relates that there is no matter on the viewing line, when in fact there is. This is useful to our problem: sometimes silhouette extraction fails locally. Accounting for this uncertainty gives our model a chance to still recover the correct voxel information thanks to contributions of other images.

Let us now consider the case where our voxel is known to be empty ($\mathcal{G}_X = 0$):

$$\begin{aligned} p(\mathcal{F}_p^i \,|\, [\mathcal{G}_X{=}0], \tau) = p(\mathcal{S}{=}0 \,|\, \tau)\; \mathcal{U}(\mathcal{F}_p^i) \\ + p(\mathcal{S}{=}1 \,|\, \tau)\; \big[\; p(\mathcal{R}{=}1 \,|\, \tau)\; \mathcal{P}_d(\mathcal{F}_p^i) \\ + p(\mathcal{R}{=}0 \,|\, \tau)\; \mathcal{P}_f(\mathcal{F}_p^i) \;\big] \end{aligned} \tag{3}$$

Still, no knowledge can be inferred about detection when the voxel is not on the viewing line of p ($\mathcal{S}{=}0$). Yet in the case where voxel X is on p's viewing line ($\mathcal{S}{=}1$), we cannot yet draw conclusions about its detection state. By definition, $\mathcal{R}{=}1$ accounts for the possibility that some other object lies on the same viewing line as the voxel: in this case detection is again ruled by the distribution $\mathcal{P}_d(\mathcal{F}_p^i)$. However, in the case no other object obstructs the viewing line ($\mathcal{R}{=}0$), detection is ruled by distribution $\mathcal{P}_f(\mathcal{F}_p^i)$. We set this distribution using a constant $P_{FA} \in [0,1]$, a parameter of our system: $\mathcal{P}_f([\mathcal{F}_p^i = 1]) = P_{FA}$ is the false alarm rate of a pixel sensor. False alarms occur when the sensor falsely relates the presence of matter on its viewing line, when in fact there is none. $\mathcal{P}_f([\mathcal{F}_p^i = 0]) = 1 - P_{FA}$ is the rate with which we expect this pixel to correctly report non-detection.

We must assign a parametric form to $p(\mathcal{R} \,|\, \tau)$. There can be detection causes anywhere along the viewing line of p. We make no assumption about these causes and consider that detection is equally likely to be triggered by the voxel occupancy or by these causes. We therefore set this term to uniform. By doing this, we consider that accounting for the possibility itself is what is important, without necessarily giving an elaborate form to this term.

Parametric form for Sampling Term $p(\mathcal{S} \mid \tau)$. This term is dependent on i, p and X. We use uniform sampling, with $p(\mathcal{S} \mid \tau) = \mathcal{U}_{k\times k}(x - p)$. This gives equal weight to all voxels that fall within a $k \times k$ window around pixel p. A smoother, normal-based sampling could also be used but requires a higher computational cost to integrate information. Generally, the shape of this sampling function can easily be modified for specific needs. Both uniform and normal sampling forms enable some control over calibration, mis-synchronization, and some classification errors: several pixels will be able to contribute to a single voxel's decision upon inference. Thanks to the introduction of these two hidden processes and the given parametric forms, our method unifies broad silhouette uncertainty management and simple image sampling methods used in some visual hull algorithms such as [2]. It also enables to embed sub-voxel information about the underlying shape in the probability grid, as opposed to purely discrete approaches such as [14].

3.3. Image Formation Term

The image pixel likelihood term $p(\mathcal{I}_p^i \mid \mathcal{F}_p^i, \mathcal{B}_p^i, \tau)$ explains the color information of a pixel (i,p), given the knowledge of the background color and silhouette detection state at this pixel. We give two parametric forms to this term. If an object detection occurred at pixel (i,p), the knowledge about background images is irrelevant to the pixel's expected color: the background is known to be occluded by an object of interest, whose color the pixel observes. With no further assumption about colors of objects of interest, we consider them uniformly distributed: $p(\mathcal{I}_p^i \,|\, [\mathcal{F}_p^i = 1], \mathcal{B}_p^i, \tau) = \mathcal{U}(\mathcal{I}_p^i)$. Reciprocally, if no object detection occurred at this pixel, then the pixel's observed color should look similar to the pixel's background color. Such an expectancy can easily be formulated using a classical background model [17]:
$p(\mathcal{I}_p^i \mid [\mathcal{F}_p^i{=}0], [\mathcal{B}_p^i{=}(\mu_p^i, \sigma_p^i)], \tau) = \mathcal{N}(\mathcal{I}_p^i \mid \mu_p^i, \sigma_p^i)$, where (μ_p^i, σ_p^i) are the parameters of a Gaussian. The method could easily use any other background model, such as a

mixture of Gaussians [15], for sub-pixel noise robustness. Nevertheless, some problems persist whatever the background model: color ambiguities between foreground and background objects, lighting, or scene geometry change. It is the goal of our integrated multi-view approach to compensate for these weaknesses of single-view estimation.

4. Voxel Occupancy Inference

Once the joint probability distribution has been fully determined, it is possible to use Bayes' rule to infer the probability distributions of our *searched* variable $\mathcal{G}_X$, given the value of our *known* variables $\mathcal{I}, \mathcal{B}, \tau$, and marginalizing over *unknown* variables $\mathcal{F}$:

$$p(\mathcal{G}_X \,|\, \mathcal{I}, \mathcal{B}, \tau) = \frac{\sum_{\mathcal{F}} p(\mathcal{G}_X, \mathcal{I}, \mathcal{B}, \mathcal{F}, \tau)}{\sum_{\mathcal{G}_X, \mathcal{F}} p(\mathcal{G}_X, \mathcal{I}, \mathcal{B}, \mathcal{F}, \tau)}$$
$$= \frac{\prod_{i,p} \sum_{\mathcal{F}_p^i} p(\mathcal{F}_p^i \,|\, \mathcal{G}_X, \tau) p(\mathcal{I}_p^i \,|\, \mathcal{F}_p^i, \mathcal{B}_p^i, \tau)}{\sum_{\mathcal{G}_X} \prod_{i,p} \sum_{\mathcal{F}_p^i} p(\mathcal{F}_p^i \,|\, \mathcal{G}_X, \tau) p(\mathcal{I}_p^i \,|\, \mathcal{F}_p^i, \mathcal{B}_p^i, \tau)} \qquad (4)$$

after substitution of (1), and factorization. More details can be found in [6].

Note that the final inference expression (4) deceptively relates our voxel occupancy to *all* pixel observations. As we compute this inference *per voxel*, this is of course intractable. In practice, detection probabilities of pixels too far from the voxel's projection degenerate to uniform, as expressed in equations (2) and (3). Their contribution therefore factors out of the inference expression (4). The inference product can then be computed over a $k \times k$ window of pixels centered at the image projection of X, in each image. With a voxel grid size of N^3, the complexity of inferring all voxels of the grid is then $O(n\ k^2 N^3)$.

5. Results and Applications

We have implemented the proposed fusion approach, using uniform voxel sampling for experiments. Compared to normal sampling it is a good trade-off between computational cost and power of information integration. Notably the method has only three parameters $\{P_D, P_{FA}, k\}$, respectively the detection and false alarm rates, and the sampling window size, all of which can often be fixed for a given application. P_D and P_{FA} ponderate the confidence given to the observations. If $P_{FA}=0$ and $P_D=1$, then we trust observations blindly. If P_{FA} and P_D are close to 0.5 then observations are not trustworthy: it takes many more observations to conclude about the occupancy. k decides how broadly each image is sampled. We have tested the algorithm under various conditions, as it can be applied to many application fields. An associated video of results is available[1].

[1] http://movi.inrialpes.fr/Publications/2005/FB05/SilhouetteCueFusion.avi

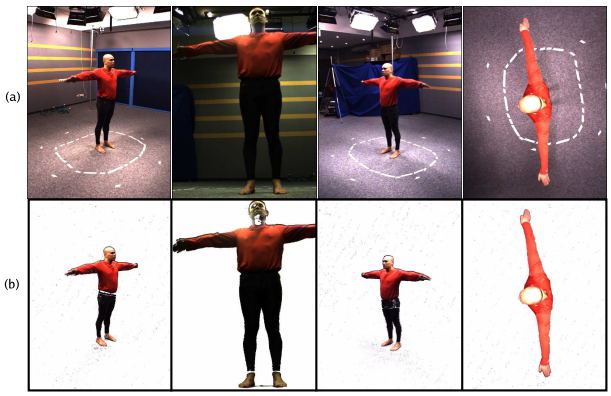

Figure 3. Inputs. (a) Four of the eight input images of the walking sequence (8 cameras, 15Hz acquisition) (b). Result given by monocular subtraction (semi-transparent rendering pondered by silhouette probability). Difficulties: camera 2 misses the subject's left forearm. Holes and noise appear in various silhouettes.

Modeling from Images. The grid itself is an estimate of shape. We illustrate this using the walking sequence. This sequence was acquired using 8 cameras of different characteristics (640×480, 780×580) at 15Hz. As Fig. 3 illustrates, the silhouette information that can be retrieved using monocular background subtraction is noisy. Also note that some cameras may not see the entire object during the sequence. These single-view subtractions also use a Gaussian background model, and reflect what input is available to our algorithm. Fig. 4 shows our method's results on a frame of the walking sequence, using a 120^3 grid. Cross-sections show how the shape information is embedded in the grid. See the associated video[1] for a dynamic view. As shown in Fig. 4(c), good surface modeling results can be achieved by extracting an isosurface from the probability grid. Fine, sub-voxel detail of the surface is recovered, and holes occurring in monocular subtractions are often filled. Additional modeling results are shown in Fig. 5.

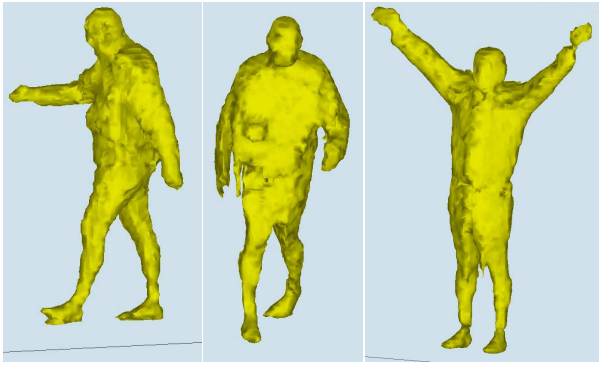

Figure 5. Isosurface of probability 0.80 at different time instants of the walking sequence. See video[1].

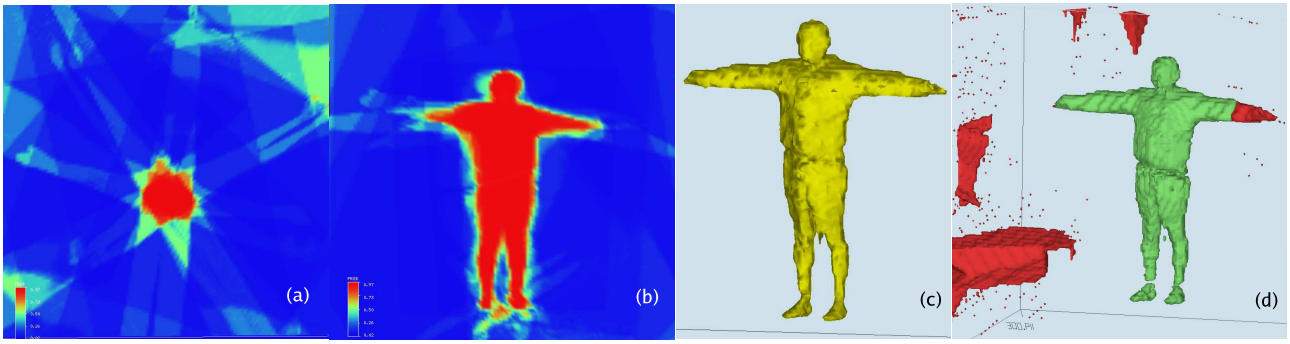

Figure 4. Walking sequence, acquired at 15Hz, using 8 cameras, with a 120^3 voxel grid. Computation time: approximately 13s on a 2.4 GHz PC. Parameters used: $P_D = 0.9, P_{FA} = 0.1, k = 5$ (a) Horizontal (chest) cross-section of the grid. Upper-left greenish regions are not seen by any camera (probability 0.5). (b) vertical grid cross-section. (c) Isosurface of probability 0.80 obtained from the grid. (d) Two classical visual hull reconstruction schemes: in light color, assuming common visibility of the object by all cameras. The forearm is lost. In dark, assuming that what is outside the visibility domain of a camera can be part of the visual hull. The latter recovers the left forearm, but ghost objects appear, in regions located in the visibility domain of a small number of cameras. Ghost objects appear when such regions project inside all silhouettes of views where they are seen.

The classical voxel-based visual hull approach has been implemented for comparison, with results in Fig. 4(d), where each voxel is carved if it projects outside silhouettes. We use the background subtractions of Fig. 3 for this experiment, and manually choose the best threshold *in each image* to provide binary silhouettes to the algorithm. Some holes are left unfilled by this method. Note that our method recovers valid occupancies from views that don't see the object entirely. This is transparent to the algorithm, because it only integrates information from sensors which see the voxels. This is unlike all classical, surface or volumetric visual hull approaches, where explicit assumptions are made about regions that project outside the visibility domain of an image, with various implications (see Fig. 4(d)).

Multi-View Background Subtraction. Our method computes a fusion of silhouette cues. This information can be used to compute consistent silhouettes in our input images, by re-projecting and rendering the occupancy grid from our input views, using a maximum intensity projection approach (MIP): for each pixel in an image, we collect the maximum probability in the grid along its viewing line. The goal is to express where silhouette detection is more likely in images. It would be possible to use the proposed statistical model to infer silhouette probability maps, given all image observations. This is however very expensive as it requires marginalization over voxel states, thus the proposed heuristic for multi-view background subtraction. All silhouettes can be extracted using a single threshold for all images. The advantage over monocular silhouette extraction is that each view benefits from the knowledge of silhouette information in the other views: resulting silhouettes show improvement, with fine details preserved (see Fig. 6). Small aliasing artifacts may appear depending on grid resolution and scene configuration.

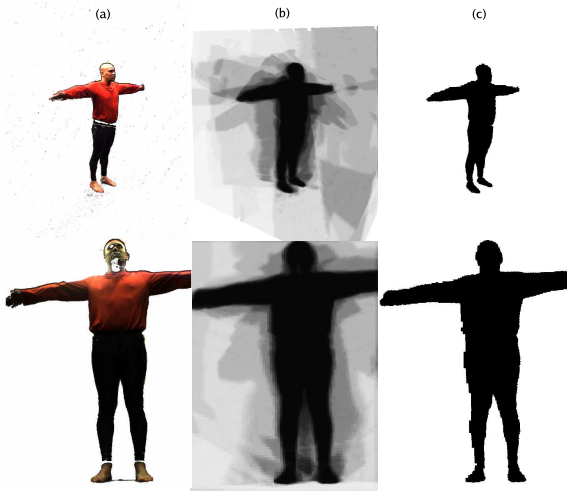

Figure 6. Multi-view silhouette extraction. (a) Monocular subtraction. Note the various artifacts and holes in such silhouettes (waist, head, feet). (b) MIP rendering of occupancy grid (120^3) probabilities from original viewpoints. Darker regions are more likely to be silhouette regions. (c) Thresholded version of (b), using a common threshold. Silhouettes show improvement. Unwanted dilatation only appears in concave regions, seen empty by a small number of cameras (crotch): the method outperforms most low-level monocular silhouette repairing schemes, such as morphological operations, for such large artifacts.

Object detection. The method can be used in much harder conditions to infer scene information. In the presence of high levels of noise, the size of the sampling window can be increased for additional robustness, with however a negative impact on precision (this tends to dilate the probability volume). Such noisy conditions limit the use of the method for 3D modeling and precise surface extraction; however the method can still be used reliably to *locate* objects in the scene. We illustrate this potential for object

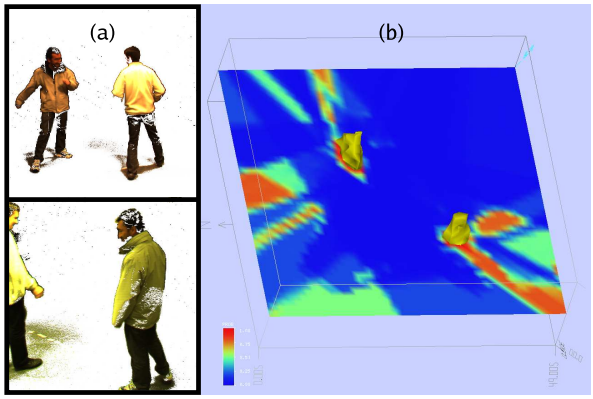

Figure 7. Multi-object sequence, with 8 cameras. (a) Difficult conditions yield very noisy single-view silhouette extractions. (b) Coarse grid ($50 \times 50 \times 18$) of the scene reconstructed with our method (computation time 7s), sufficient to localize objects (using $k = 25$). A horizontal cross-section of the grid, as well as the 0.67-probability isosurface, are shown (see the associated video). This is sufficient to localize the people.

localization, in an experiment with loose camera configurations and poor contrast images (see Fig. 7). 8 cameras are placed such that a relatively large area ($25m^2$) can be monitored in the room. Most cameras see the center of the room, but peripheral regions of this area are seen by 3 or 4 cameras at most. Two people walk randomly in the scene and are successfully localized, when seen by at least 3 cameras.

6. Discussion

We have presented a novel approach for silhouette cue fusion from multiple views. We use a rigorous sensor fusion framework, to relate scene information directly to observations. This has various advantages: the entire causality chain is modeled and all assumptions made explicit. It also avoids making hard decisions about silhouette labeling in images, which would have required tedious per-image parameter settings. Thus the underlying silhouette information in images can be smoothly integrated, using only three global parameters of a pixel sensor model. These parameters intuitively express the reliability of observations. This approach has been validated with several applications, and many new ideas can be experimented and plugged-in without changing the core of the method.

Arguably, more dependencies could be considered in the model. Namely, we notice that the reliability of pixel decision can be related to the colors observed at this pixel: many times we observe the case where black foreground objects are observed in front of a black background and misclassified, a case which could be explicitly modeled. The local nature of grid evaluations opens the possibility for a real-time, hardware-accelerated solution. More generally, our model estimates static grids at one time instant. It would greatly benefit from temporal consistency, where passed observations are used to infer current occupancy states. Happily, occupancy grids provide a good framework for temporal accumulation of information, being one of its main uses in the robotics community [3]. We will investigate these possibilities to extend the capabilities of our system.

References

[1] A. Broadhurst, T. Drummond, and R. Cipolla. A probabilistic framework for the Space Carving algorithm. In *ICCV'01*, pages 388–393, 2001.

[2] K. M. Cheung, T. Kanade, J.-Y. Bouguet, and M. Holler. A real time system for robust 3d voxel reconstruction of human motions. In *CVPR'00*, volume 2, pages 714 – 720, 2000.

[3] A. Elfes. *Occupancy grids: a probabilistic framework for robot perception and navigation*. PhD thesis, 1989.

[4] O. Faugeras and R. Keriven. Complete dense stereovision using level set methods. In *ECCV'98*, volume I, pages 379+, 1998.

[5] J. S. Franco and E. Boyer. Exact Polyhedral Visual Hulls. In *BMVC'03*, 2003.

[6] J. S. Franco and E. Boyer. Fusion of Multi-View Silhouette Cues Using an Occupancy Grid. Technical report, INRIA RR-5551, Apr. 2005.

[7] B. Goldlücke and M. Magnor. Joint 3-d reconstruction and background separation in multiple views using graph cuts. *CVPR'03*, I:683–694, 2003.

[8] K. Grauman, G. Shakhnarovich, and T. Darrell. A bayesian approach to image-based visual hull reconstruction. In *CVPR'03*, pages 187–194, 2003.

[9] J. Isidoro and S. Sclaroff. Stochastic refinement of the visual hull to satisfy photometric and silhouette consistency constraints. In *ICCV'03*, pages 1335–1342, 2003.

[10] K. Kutulakos and S. Seitz. A Theory of Shape by Space Carving. *IJCV*, 38(3):199–218, 2000.

[11] A. Laurentini. The Visual Hull Concept for Silhouette-Based Image Understanding. *IEEE Transactions on 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. In *International Conference on Machine Learning*, pages 332–340, 1998.

[13] W. Matusik, C. Buehler, R. Raskar, S. Gortler, and L. McMillan. Image Based Visual Hulls. In *ACM Computer Graphics (Proceedings Siggraph)*, pages 369–374, 2000.

[14] D. Snow, P. Viola, and R. Zabih. Exact voxel occupancy with graph cuts. In *CVPR'00*, pages 345–353, 2000.

[15] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR'99*, pages 2246–2252, 1999.

[16] R. Szeliski. Rapid Octree Construction from Image Sequences. *Computer Vision, Graphics and Image Processing*, 58(1):23–32, 1993.

[17] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on PAMI*, 19(7):780–785, 1997.

[18] G. Zeng and L. Quan. Silhouette extraction from multiple images of an unknown background. In *ACCV'04*, 2004.