Shape From Inconsistent Silhouette

Jose-Luis Landabaso   Montse Pardàs   Josep Ramon Casas

Jose-Luis Landabaso

Telefonica I+D (TID)

Via Augusta 177, 08021 Barcelona, Spain

e-mail: jlldiaz@tid.es

Tel: +34649192778, Fax: +34934016447

Montse Pardàs (corresponding author)

Technical University of Catalunya (UPC)

Jordi Girona 1-3, 08034 Barcelona, Spain

e-mail: montse@gps.tsc.upc.edu

Tel: +34934015894, Fax: +34934016447

Josep Ramon Casas

Technical University of Catalunya (UPC)

Jordi Girona 1-3, 08034 Barcelona, Spain

e-mail: josep@gps.tsc.upc.edu

Tel: +34934016458, Fax: +34934016447

Abstract

Shape from Silhouette (SfS) is the general term used to refer to the techniques that obtain a volume estimate from a set of binary images. In a first step, a number of images are taken from different positions around the scene of interest. Later, each image is segmented to produce binary masks, also called silhouettes, to delimit the objects of interest. Finally, the volume estimate is obtained as the maximal one which yields the silhouettes. The set of silhouettes is usually considered to be consistent which means that there exists at least one volume which completely explains them. However, silhouettes are normally inconsistent due to inaccurate calibration or erroneous silhouette extraction techniques. In spite of that, SfS techniques reconstruct only that part of the volume which projects consistently in all the silhouettes, leaving the rest unreconstructed. In this paper, we extend the idea of SfS to be used with sets of inconsistent silhouettes. We propose a fast technique for estimating that part of the volume which projects inconsistently and propose a criteria for classifying it by minimizing the probability of miss-classification taking into account the 2D error detection probabilities of the silhouettes. A number of theoretical and empirical results are given, showing that the proposed method reduces the reconstruction error.

Keywords

Shape from Silhouette, multi-camera background subtraction, Visual Hull

List of Figures

1 Introduction

Shape extraction from a set of silhouettes (binary masks of the objects of interest in the foreground scene) was firstly introduced by Baugmart [1] in 1974, though it was not until 1991 when Laurentini [2] defined the geometric concept of Visual Hull (VH) as the maximal object silhouette-equivalent to the real object S, i.e., which can be substituted for S without affecting any silhouette [3,4,5]. Since then, Shape from Silhouette (SfS) has been considered as the method of obtaining the VH of an object.

The concept of VH is strongly linked to the one of silhouettes' consistency: A set of silhouettes is consistent if there exists at least one volume which exactly explains the complete set of silhouettes, and the VH is the maximal volume among the possible ones. If the silhouettes are not consistent, then it does not exist an object silhouette-equivalent, that is, the VH does not exist. Total consistency hardly ever happens in realistic scenarios due to inaccurate calibration or noisy silhouettes caused by errors during the 2D detection process: background learning techniques [6,7,8,9,10,11,12,13,14], chroma key techniques [15,16], etc. In spite of that, most SfS methods have been designed in the past assuming that the silhouettes are consistent, thus reconstructing only the part of the volume which projects consistently in all the silhouettes, i.e., the volume where the visual cones intersect, without further considerations.

We propose a shape reconstruction method based on the silhouette consistency principle. Our system validates the regions in the silhouettes which are consistent in all the projections and adjusts the regions which are not, dealing with 2D errors, i.e., misses (foreground voxels detected as background) and false alarms (background voxels detected as foreground), in an unbiased way. By contrast, other SfS systems usually treat differently the 2D errors on the basis of their type.

In the following, we summarize the different techniques available for extracting shapes from a set of silhouettes. Then, we discuss which are the different types of 2D errors and how they affect the reconstructed shape.

1.1 Shape from Silhouette

Many algorithms have been developed for constructing volumetric models from a set of silhouette images. Silhouette images are first extracted by creating statistical models of the background process of every pixel value, i.e. color [11,6,10], texture [17,18,19], or temporal-based information [20]. Then, the foreground segmentation is performed at each pixel, either as an exception to the modeled background [6,7,8,9,10,11], or in a Bayesian framework, using a maximum a posteriori classifier if there exist foreground models [12,13,14]. A good review of the state-of-the art on 2D-background subtraction methods can be found here [21]. Once the silhouettes are extracted, the main step of all the algorithms is the intersection test. Some methods back-project the silhouettes, creating an explicit set of cones that are then intersected in 3D [22,23,24]. Others divide the volume into voxels [25,26,27,28,29,30]. Then each voxel is projected into all the images to test (using a projection test) whether they are contained in every silhouette. More efficient octree-based strategies have also been used to test voxels in a coarse to fine hierarchy [31,32]. See [33,34] for two surveys on volumetric-based methods.

Accurate silhouette extraction is crucial for good performance of SfS, independently of the algorithm used. Following, we discuss how errors in the silhouettes affect the reconstructed Shape. Based on the outcomes of the issues discussed, a more in-depth analysis of the proposed 3D-reconstruction technique will be possible.

1.2 Noise Propagation to the 3rd Dimension

Silhouette image noise can be classified in different ways, e.g., according to the observable effects over the silhouettes; or depending on the cause that produced the error:

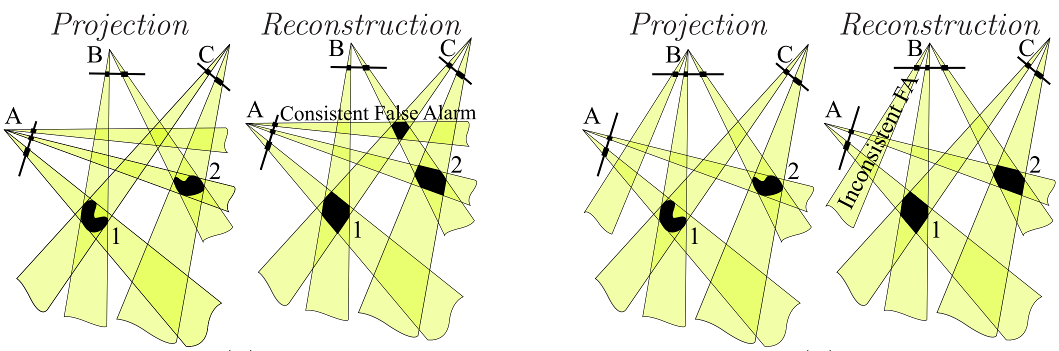
- Defects observable in the silhouettes can be categorized into two types: false alarms and misses. False alarms correspond to erroneous foreground detections, while misses correspond to erroneous background detections.
- Errors in the silhouettes can be due to different causes: regular noise and non-Gaussian deterministic errors. The first type of error is because of the cameras thermal noise. Examples are the isolated background pixels within the foreground silhouettes and foreground pixels within the background that can be observed in Figure 1. The second one often consists in large regions missed or falsely detected due to the arrangement of the scene or limitations of the foreground segmentation technique. Deterministic misses in a view often occur when, for instance, foreground objects have similar colors and texture to their counterparts in the background. Deterministic misses can also be due to background structures, such as the table in Figure 1, occluding the foreground objects in some views. Analogously, specular reflections can form large areas of falsely detected foreground pixels, see for instance the upper left corner in Figure 1.

Most techniques in the literature have been focused on reducing the effects of the Gaussian nature errors. However, since both of them can produce the

Figure 1. Original image and segmented silhouette.

same effects (false alarms and misses), the study of 3D error propagation can be isolated from the cause.

In classical SfS, a false alarm in a view does not contribute to a false alarm in 3D unless the visual cone that is erroneously created intersects simultaneously with other $C-1$ visual cones, where C is the total number of cameras (see Figure 2(a)). If the intersection is produced, then the volumetric points corresponding to the intersection are wrongly reconstructed. Since the reconstructed shape is consistent because its projection in all the views matches with the silhouettes, then the 2D false alarm is undetectable. However, the shape is not reconstructed in the parts of the volume where at least one of the erroneous visual cones does not intersect simultaneously with other $C-1$ visual cones (see Figure 2(b)). This is the most typical case in scenarios where the major part of the volume is unoccupied. In such case, the cones produced by 2D false alarms do not intersect with visual cones from the rest of cameras, then 2D false alarms are inconsistent with the reconstructed shape, allowing their detection as we will show in the following sections.



(a) (b)

Figure 2. In (a) there has been a false detection in camera A. The false visual cone intersects with other $C-1$ visual cones forming a false shape reconstruction. Another false alarm in camera B is depicted in (b). In this case, the false alarm forms an inconsistent cone for not intersecting with other $C-1$ visual cones. This type of false alarm, which is the most common case, does not affect SfS reconstructions.

Contrarily, a miss in a view inhibits the simultaneous intersection of C visual cones in 3D, leading to an ineluctable miss in the shape (see Figure 3). This makes the SfS algorithm highly sensitive to this type of errors, whereas 2D false alarms do not produce erroneous reconstructions in most of the cases. 2D misses can also be indirectly detectable, since the projection of the incomplete Visual Hull reconstructed will not match with the rest of correct silhouettes.

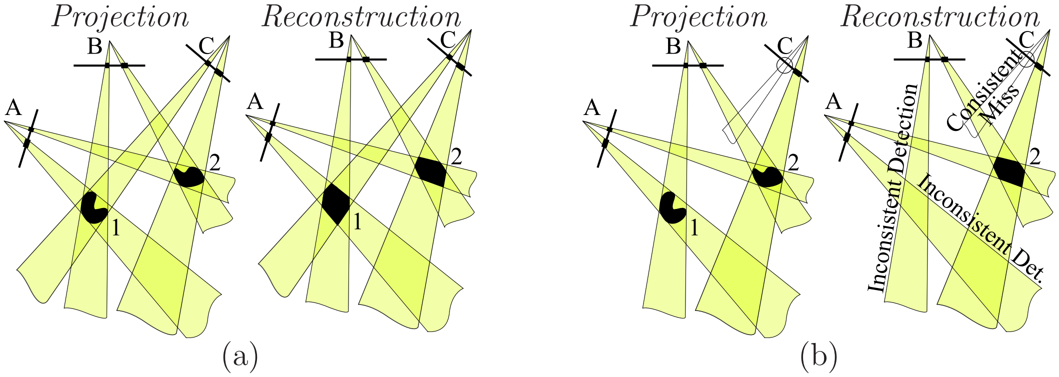


Figure 3. In (a), objects 1 and 2 are correctly detected in all the cameras. In (b), object 1 has been missed in camera C. On the right the Visual Hull is depicted. Note that the visual cones which do not intersect with any reconstructed shape are considered to be inconsistent with respect to the Visual Hull.

As a final thought on the effects of 2D error propagation, it seems clear that the very sensitive response of SfS to 2D misses contradicts the general notion that "as the number of cones increases, the object is reconstructed with higher precision" [3]. While this is true with perfectly extracted silhouettes, it is not the case when the silhouettes have *non null rates of miss*. In fact, an infinite number of silhouettes with a low but non null rate of randomly distributed misses will not reconstruct any shape. In conclusion, although SfS algorithms are perfectly fine with consistent silhouettes, they tend to penalize 2D misses in front of 2D false alarms when the silhouettes are inconsistent. The Shape from an Inconsistent set of Silhouettes (SfIS) has to be based on a different principle; one that takes decisions in accordance with the probabilities of 2D false alarm and miss; and one which does not imply that the Shape lies only in the intersection of *all* the visual cones.

Indeed, SfIS might introduce more false alarms to the Shape than SfS as the payoff for recovering some of the misses. We will show that false alarms will be introduced only to the extent that global error is lower than without them.

1.3 Dealing with noise in related works

In the past, efforts have been put in proposing different algorithms for palliating the effects of the propagation of the 2-dimensional noise. There are different approaches to achieve noise reduction.

The first general approach involves using voxel-based reconstructions to reduce the probability of voxel miss-classification. In [29], Cheung et al. propose an algorithm called SPOT. In their approach, the voxels are projected into each camera view. Then, their algorithm determines the minimum number of foreground pixels (Z_ϵ) which have to be detected inside each projection of a voxel to consider that the projection test is passed in a certain view. Finally, if the projection test is passed *in all the views*, then the voxel is classified as part of the Shape. The minimum number of foreground pixels Z_ϵ, over the total Z, is determined after minimizing the probability of voxel miss-classification considering that the silhouettes are *consistent* (i.e., that a voxel is part of the shape if and only if the projection test is passed in all the views, while it is background otherwise). So, on the one hand, SPOT considers that the masks are consistent while, on the other hand, it accepts that the masks are inconsistent for having misses and false alarms. Under the same assumption of silhouette consistency, SPOT achieves lower voxel miss-classification rate compared to other SfS algorithms that use naive projection tests such as testing only one point per voxel and view or testing all the pixels within the projection of the voxel. But even though SPOT and other voxel-based noise reduction methods are an important step forward, none of them have focused on the detection of deterministic errors.

A second approach suggested in [28] as a reference for comparison with the method proposed, and also used in [27], requires the intersection of at least $C-P$ visual cones to allow a reconstruction, where P is the number of acceptable misses among the set C of cameras. Although single misses do not block the reconstruction in this approach, the resulting shape is larger than the real Visual Hull for requiring fewer intersections of visual cones. A drawback of this approach is that larger hulls are reconstructed either if the silhouettes are consistent or not.

Another approach is to classify voxels as shape or background using cooperatively the information from the multiple cameras before extracting the 2D silhouettes. In [28] an algorithm based on graph cuts determines the 3D shape with lowest cost (smoothest shape consistent with the observations). In this case, the 2D silhouettes are not explicitly computed. In [35] the shape-from-silhouette problem is restated as a sensor fusion problem, providing each pixel from each camera with a forward sensor formulation which models the pixel observation responses to the voxel occupancies in the scene. Finally, in [30] the classification of the 3D space is made on a Bayesian framework, using the 2D foreground/background probabilities of the multiple views. Although these approaches are more robust to image noise and calibration errors than standard Shape from Silhouette, they do not consider deterministic errors caused by occlusions or failures of the foreground detection technique. In this paper, we propose to take advantage of the inconsistencies between the reconstructed shape and the silhouettes to further improve the resulting shape.

Multi camera consistency constraints provide tools for detecting deterministic errors. This has been used in another context in [36], where silhouette consistency is used to determine the relationship between sets of silhouettes captured with an object in a different pose corresponding to each silhouette. The epipolar tangency constraint (testing correspondences of the frontier points) is used as a necessary condition for shape consistency. The authors discard using the area of each silhouette that lies outside the visual hull for being slow in this context and not suitable for pose estimation [36].

Our approach is placed in the later context. We propose a fast technique for estimating that part of the volume which projects inconsistently and propose a criteria for classifying it either as part of the shape or not by minimizing the probability of voxel miss-classification. Our approach is voxel-based and can be used to correct errors from any Shape from Silhouette technique, from the standard ones to those which were proposed to minimize the effects of noise in the foreground detection [28,35,30]. Moreover, we propose a general framework where any projection test can be used [29].

The proposed method reconstructs the VH with standard Shape from Silhouette in the first step. In silhouette-based systems, Shape from Silhouette reconstructs the volume with the lowest classification error for those voxels that project consistently to all the camera views. A decision on the voxels not forming part of the VH is taken in a second step by minimizing the error probability on each voxel independently. In order to compute this error probability the projection of the computed VH is compared to the set of original silhouettes.

The remainder of the paper is structured as follows. In the next section, the voxel-based SfS approach is discussed. Section 3 is devoted to discussion of SfIS, including detailed algorithms for its implementation. Section 4 presents the conditions in which a very fast implementation of SfIS is possible. In section 5, theoretical and experimental studies of the system are presented with various synthetic and real-world test images. Finally, the paper concludes in section 6.

2 Voxel-Based Shape from Silhouette

In SfIS, volume classification is achieved after minimization of the probability of miss-classification. Since 3D errors depend on the classification probability of the 2D technique, it is important to first study which are these 2D-error probabilities. To do so, we focus on the voxel-based approach and discuss the probabilities of error of several projection tests. In addition, we give the probability of error of the voxel-based SfS approach, so that we can compare it later with the probability of error of SfIS.

First off, the voxel-based SfS algorithm for any projection test is the one shown in Algorithm 1.

Require: Silhouettes: $S(c)$, a Projection Test Function: $PT_c(voxel, Silhouette)$

```
for all voxel do
    voxel ← Foreground
    for all c do
        if PT_c(voxel, S(c)) is false then
            voxel ← Background
```

Algorithm 1: Voxel-based SfS algorithm

Since voxel classification errors may be due to either false alarms or misses, the probability that a voxel is miss-classified is:

$$P(Err_{3D}) = P_B P(FA_{3D}) + P_S P(M_{3D}), \tag{1}$$

where P_B and P_S are prior probabilities of a voxel forming part of the Background or Shape, respectively [1], and $P(FA_{3D})$ and $P(M_{3D})$ correspond to the probabilities of false alarm and miss in a voxel.

Since 3D false alarms in classical SfS happen when a voxel is wrongly classified as part of the Shape in all camera views, while misses happen when a voxel is wrongly classified as part of the Background in at least one camera view:

$$P(Err_{3D}) = P_B \underbrace{\prod_{i=1}^{C} P_i(FA_{2D})}_{P(FA_{3D})} + P_S \underbrace{\left(1 - \prod_{i=1}^{C}(1 - P_i(M_{2D}))\right)}_{P(M_{3D})}, \tag{2}$$

where $P_i(FA_{2D})$ and $P_i(M_{2D})$ correspond to the probabilities that the projection test has been wrongly passed (false alarm) or wrongly failed (miss) in camera i, respectively.

Equation (2) can be expressed more compactly when the probabilities of false alarm and miss are equiprobable in all the views ($P_i(M_{2D}) = P_j(M_{2D})$ and $P_i(FA_{2D}) = P_j(FA_{2D})$, for all i and j):

$$P(Err_{3D}) = P_B \underbrace{P(FA_{2D})^C}_{P(FA_{3D})} + P_S \underbrace{\left(1 - (1 - P(M_{2D}))^C\right)}_{P(M_{3D})} \tag{3}$$

that besides being more compact, it is also significantly faster to compute than equation (2). In the following, we will refer to a test as equiprobable if it has

[1] Priors P_S and $P_B = 1 - P_S$ can be simply obtained by computing the detected/total voxel occupancy ratio, for instance.

equal error probability in all views, and non-equiprobable if it has different error probability in each view where the test is being carried out.

Different projection tests can be used. The most simple one being the Single Pixel Projection Test that projects the point in the center of a voxel into a pixel in all the camera views. A Complete Pixels Projection Test can also be used, consisting in testing all the pixels within the splat of the voxel in each camera. The test is passed only if all the pixels within the splat belong to the silhouette, that is, they are foreground pixels. The Incomplete Pixel Projection Test can be defined as a soft version of the Complete Test, in which the test is passed in a in a view i when a minimum number of pixels M_i over all pixels belonging to the splat (S_i) lie in the silhouette. A more efficient and robust projection test is the Sampled Pixels Projection Test, that we define in the following.

2.1 Sampled Pixels Projection Test

We have developed the Sampled Pixels Projection Test with SPOT [29] as principal inspiration. As with SPOT, a number of R points within the voxel are selected. These points may be equidistant among them, or just randomly selected. The test is passed in a view i when at least N projected points, i.e. pixels, over the total R, are within silhouette i.

$$\begin{aligned} &\text{Pixels in the Shape} \geq N \Rightarrow \text{pass the test} \\ &\text{Pixels in the Shape} < N \Rightarrow \text{do Not pass the test} \end{aligned} \tag{4}$$

Selection of R points for each voxel makes the test very fast for two reasons:

The first reason is that the number of selected points is chosen independently of the voxel position, and therefore the probabilities of voxel miss-classification are the same for all voxels.

The second reason is that since the test is run using exactly R pixels in each projection, the probabilities of false alarm and miss of the test are identical in all views. Thus, one needs to compute the probability of miss-classification of a projection test only once for all views. Furthermore, as a consequence of error equiprobability in the views, $P(Err_{3D})$ can be computed using faster equation (3), instead of (2).

The Sampled Pixels Projection Test here proposed differs from SPOT in the expressions used to calculate the probability of voxel miss-classification. In SPOT it is assumed that priors P_B and P_S are equiprobable, which is almost never the case, being in some setups P_B several orders of magnitude larger than P_S. Another difference is that SPOT considers that a voxel-miss occurs only when *exactly* one projection test is wrongly failed which is computationally less complex, while the Sampled Pixels Projection Test uses equation (3) which considers voxel-misses when the projection test is missed in *at least* one view (see equation (5)):

$$P(M_{3D}) = 1 - (1 - P_i(M_{2D}))^C = \sum_{i=1}^{C} \binom{C}{i} P(M_{2D})^i (1 - P(M_{2D}))^{C-i} \quad (5)$$

In order to fully express equation (3), the probabilities of false alarm ($P(FA_{2D})$) and miss ($P(M_{2D})$) of the test have to be deduced. In the Sampled Pixels Projection Test, these probabilities depend on N in the following manner:

Since the test is passed when at least N pixels lie in the silhouette, false alarms of the projection test happen when there are at least N pixels falsely detected. Contrarily, misses of the projection test occur when there are at least $R-N+1$ pixels missed.

Based on this reasoning, both miss-classification probabilities have to add together the probabilities of all the possible cases which lead to a miss-classification. Table 1 shows the precise mathematical expressions of $P(FA_{2D})$ and $P(M_{2D})$ of the test.

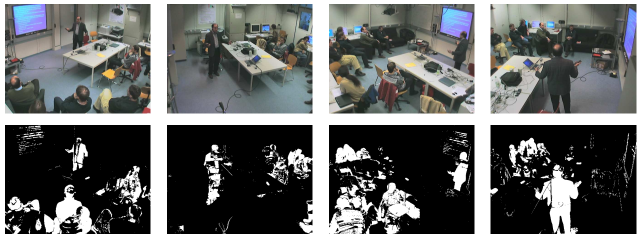
Once that $P(Err_{3D}[N])$ has been expressed, the following step is to choose the minimum number of points N over R which have to belong to the silhouette so that the test is passed. Indeed, the best N is the one which minimizes the probability of voxel miss-classification:

$$N^{\star} = \operatorname*{argmin}_{N} P(Err_{3D}[N]) \quad (6)$$

Since $P(Err_{3D}[N])$ is not continuous, it cannot be minimized by differentiating it. However, the optimal $N^{\star}$ can be obtained by doing an exhaustive search over all possible $N \in [0, R]$. Note that even though being computationally demanding, the calculation does not entail a problem since it only has to be performed once for all views and voxels.

In Figure 4 [2] we depict some visual examples of the behavior of the presented projection tests. Note that Single Pixel Test is not as good performer as the

[2] The images correspond to the 2005 evaluation dataset used within the framework of the CHIL *Computers in the Human Interaction Loop* project [37]. The images were acquired in the smart-room of the Interactive Systems Labs at the University of Karlsruhe, Germany. The setup includes 4 fully calibrated wide angle lens cameras with a resolution of 768×576 pixels, positioned at the room corners.



Original images and silhouettes extracted using [6].

VH projection using SfS with the Single Pixel Projection Test.

VH projection using SfS with the Sampled Pixels Projection Test.

Figure 4. All the VHs have been reconstructed in the area of the presenter, using voxels with edge size of 2.5 *cm*.

Sampled Pixels Test, in this example. However this approach is suitable for low-error systems focused to real-time operation.

In Table 1 we present a summary of the error probabilities for these two projection tests, which can be easily derived from their definition. Both projection tests can be combined with the standard voxel-based algorithm. Following, we propose the SfIS algorithm, which also makes use of the proposed tests. In SfIS, the error probabilities of the employed test are important, and therefore Table 1 will be a useful resource for the implementation of the algorithm.

Table 1
Projection Tests Error Probabilities

Type	Proj. Test	Error Probability
$F.A.$: $P(FA_{2D})$	Single	$P(FA_{pix})$
	Sampled	$\sum_{i=N}^{R} \binom{R}{i} P(FA_{pix})^i (1 - P(FA_{pix}))^{R-i}$
$Miss$: $P(M_{2D})$	Single	$P(M_{pix})$
	Sampled	$\sum_{i=R-N+1}^{R} \binom{R}{i} P(M_{pix})^i (1 - P(M_{pix}))^{R-i}$

3 Shape from Inconsistent Silhouette (SfIS)

In SfIS, the VH is reconstructed using SfS methods and corrected later with those parts of the volume which were not correctly classified. 3D miss-classifications can be detected by examining the inconsistent regions of the silhouettes. To detect inconsistent regions, one can project back the VH and test whether the projections match with the generative silhouettes. Then, the shape can be reconstructed using a different criterion when there are parts of the volume (*Inconsistent Volume:IV*) which project to inconsistent regions in the silhouettes (*Inconsistent Silhouettes:ISs*). Preliminary work on SfIS was presented in [38]. In the following, we provide the generalization of SfIS for any type of projection test. First, we formalize the concepts of IV and ISs and propose a procedure for estimating them. Then, we propose a method for optimally classifying the IV into Shape or Background.

3.1 Inconsistent Volume (IV)

The geometric concept of IV is introduced as the volume where there does not exist a shape of the VH which could possibly explain the observed silhouettes. The ISs are the resulting silhouettes after subtracting the original silhouettes with the projection of the visual hull (see Figure 5 [3]).

The IV can be defined as the union of all the inconsistent cones, formed by the back-projection of the ISs into the 3D scene. Thus, when the set of silhouettes is consistent then *all* the ISs are empty, and the IV is also empty. However, when *a single* inconsistency appears in at least one silhouette then the IV will not be empty.

From the above equivalent definitions of the IV, it follows that the IV is disjoint from the VH ($VH \cap IV = \emptyset$). This can be observed in Figure 6, where different situations with consistent and inconsistent sets of silhouettes have been depicted: In (a), there are two foreground objects which are correctly detected in all the cameras. In (b), camera C misses foreground object 1. The miss-detection entails an inconsistent set of silhouettes in cameras A and B. Further inspection of the figure indicates that the IV in this case corresponds to the union of the visual cones *camA→obj1* and *camB→obj1*, confirming that is disjoint from the VH reconstructed around object 2. In (c), object 2 is correctly reconstructed but there have been two false alarms in cameras A and B. These false alarms coincide with the regions of the projection of object 1 in (b). The IV in this case is the same as in (b).

[3] The *Kung-Fu Girl* dataset is provided by the *Graphics Optics Vision group* of *Max-Planck-Institut fur Informatik*.

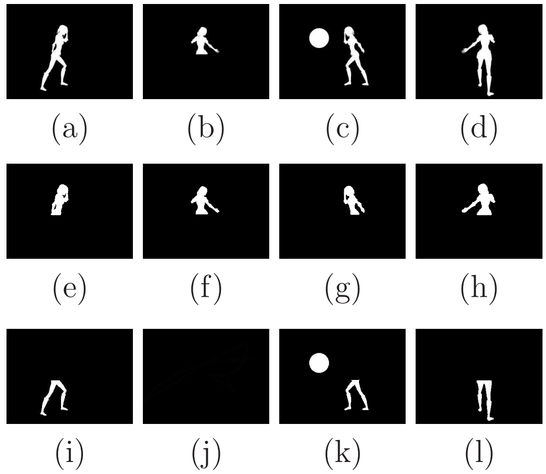
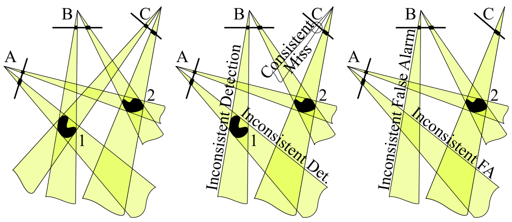


Figure 5. The first row of images shows four synthetic silhouettes, corresponding to the *Kung-Fu Girl* dataset, where some errors have been intentionally introduced: In (b), the bottom part of the silhouette has been deliberately removed and, in (c), a false alarm has been incorporated. The second row of images shows the projection of the VH reconstructed using SfS from the silhouettes above. Note that the 2D false alarm does not propagate to 3D, while a single miss propagates to 3D preventing a proper reconstruction of the VH. Finally, in the bottom row, the ISs are shown. The IV is the union of the back-projected cones of the inconsistent silhouettes.

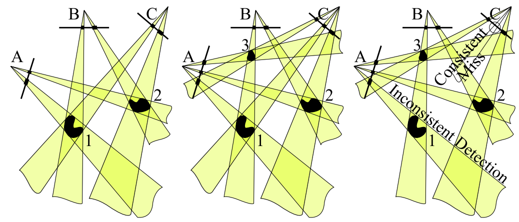
A closer look at Figure 6(b) & Figure 6(c) reveals some preliminary conclusions about how the IV could be classified. Observe that both figures depict different situations that could have been the cause of the same observed silhouettes. Note that it is impossible to guarantee whether there has been a single miss in camera C or two false alarms in cameras A and B. However, the figures suggest that the more inconsistent cones intersect, the higher the chances that the rest of cameras have missed an object in the area of inconsistent cone intersection. Of course, the exact chances of missing an object will also depend on the probabilities of 2D miss-classification. The main problem to solve will be how to choose the minimum number of inconsistent intersections ($T^{\star}$) that have to be produced so that it can be determined that a part of the Shape was missed during the reconstruction process.

There is yet another factor which will have to be considered in the choice of the optimal $T^{\star}$. As Figure 6 suggests, the number of intersecting inconsistencies is apparently tied to either the number of false alarms, or to the number of cameras minus the misses. However, there is a case for which this is not true. This situation is due to the fact that inconsistencies can be hidden by occluding objects. Figure 7 shows a typical situation with inconsistencies and occlusions. In the figure, a new object (object 3) has been deliberately placed in the same visual cone of $camB{\rightarrow}obj1$. Thus, object 3 prevents the inconsistent cone $camB{\rightarrow}obj1$ when camera C misses object 1. The figure clearly indicates that the number of inconsistent cone intersections is not a sufficient piece of information for deciding whether there have been misses in some silhouettes



(a) (b) (c)

Figure 6. In (a), objects 1 and 2 are correctly detected in all the cameras. In (b) object 2 is correctly detected in all the cameras, but object 1 is missed in camera C. Former visual cones $camB \rightarrow obj1$ and $camA \rightarrow obj1$ are now inconsistent cones, whose union forms the IV. Note that it is impossible to know, from the observed silhouettes, whether there has been a single miss in camera C or two false alarms in cameras A and B, as depicted in (c).



(a) (b) (c)

Figure 7. In (b) objects 1, 2 and 3 are correctly detected in cameras A, B and C. In (c), although objects 2 and 3 are correctly reconstructed, object 1 is not. The IV in this figure is smaller than its counterpart in Figure 6(b) due to the occlusion of object 1. The figure suggests that the number of occlusions will be an important determinant for proper classification of the IV.

or not. Furthermore, the figure also suggests that all views where an object occludes a point of the IV will have to be ignored when determining its $T^{\star}$. Note that, due to self-occlusions of the objects, occlusions are not rare, and thus they have to be considered.

In the following, we propose a method to determine the IV. Then, we describe how we choose the minimum number of inconsistent intersections that have to be produced so that it can be determined that an object was missed. The presented method will take into account previous considerations regarding occlusions.

3.2 IV voxelization

Prior to deriving the expressions for the IV classification, first we need a method to reconstruct it.

In order to estimate the IV, we need to determine the *union* of the inconsistent cones (corresponding to the back-projection of the ISs) analogously as SfS methods determine the *intersection* of the visual cones (corresponding to the back-projections of the silhouettes). As it has been previously reviewed in section 1.1, determining visual cone intersections can be performed in different ways. For instance, some SfS techniques project back the silhouettes, creating the set of visual cones which are intersected in the 3D space. In other approaches, the volume is divided in voxels which are then projected to the images to find out (using a projection test) whether they are contained in every silhouette or not.

In this paper, we develop the concept of Shape from Inconsistent Silhouette using a voxel-based approach, although similar considerations can be derived with the geometrical approach.

Require: Silhouettes: $S(c)$, Proj. Test: $PT_c(voxel, Silhouette)$
for all $voxel$ **do**
    $VH(voxel) \leftarrow$ true
    for all c **do**
        if $PT_c(voxel, S(c))$ is false **then**
            $VH(voxel) \leftarrow$ false
Project the VH to all the camera views: $VH_{proj}(c)$
for all $voxel$ **do**
    $IV(voxel) \leftarrow$ false
    for all c **do**
        if $PT_c(voxel, S(c))$ is true **then**
            if $PT_c(voxel, S(c)) \neq PT_c(voxel, VH_{proj}(c))$ **then**
                $IV(voxel) \leftarrow$ true

Algorithm 2: Voxelization of the IV

The detailed process for the IV voxelization is shown in Algorithm 2. Note that in the voxel-based approach, the role of the inconsistent silhouettes (difference between *silhouettes* and *VH projection*) is replaced by the nonequivalence of their projection tests: $PT_c(voxel, S(c)) \neq PT_c(voxel, VH_{proj}(c))$.

3.3 Unbiased Hull (UH)

The IV contains all the volumetric points which cannot justify the silhouettes where they project. In terms of *consistency*, these points are candidates of not having been classified as Shape by error, while all the points in the VH are error-free. We define the Unbiased Hull (UH) as the subset of the IV which is better explained as Shape for minimizing the probability of voxel miss-classification. We call it *unbiased*, since the volumetric points of the IV are classified as either Shape or Background based on the lowest probability of error, while VH reconstruction methods are *biased* for always classifying the IV as Background. In the final stage of the process, the union of the VH and UH will form the best apparent hull in terms of lower miss-classification probability.

The classification of the voxels in the IV has to be optimal based on all the characteristics we can gather from them:

- First off, each voxel in the IV has an associated number of foreground projections ($\mathcal{F}$), which corresponds to the number of visual cone intersections in the voxel. $\mathcal{F}$ can be simply calculated by counting the number of silhouettes ($S(c)$) where the projection test is passed: $PT_c(voxel, S(c)) = true$, being $voxel \in$ IV. For example, in Figure 7(c), all voxels corresponding to object 1 have $\mathcal{F} = 2$, for being in the visual cones *camA→obj1* and *camB→obj3*.

  In the IV, the number of foreground projections is bounded by: $1 \leq \mathcal{F} \leq C - 1$, since 0 foreground projections would correspond to a consistent background detection in the VH and C detections would correspond to a consistent foreground detection in the VH.
- A voxel in the IV can also be characterized by the number of consistent foreground projections ($\mathcal{O}$), corresponding to the number of views where the voxel has been occluded. $\mathcal{O}$ can be computed as the number of times that the projection test is passed both in a silhouette ($S(c)$) and in the projection of the VH ($VH_{proj}(c)$): $PT_c(voxel, S(c)) = true = PT_c(voxel, VH_{proj}(c))$, being $voxel \in$ IV. For instance, voxels corresponding to object 1 in Figure 7(c) have $\mathcal{O} = 1$, for intersecting with the consistent occluding cone: *camB→obj3*.

  The number of occlusions in the IV is bounded by $0 \leq \mathcal{O} \leq C - 1$, as C occlusions would correspond to a foreground detection in the VH.
- A voxel in the IV also has an associated number of inconsistencies ($\mathcal{I}$), which corresponds to the number of inconsistent foreground projections. Note that $\mathcal{I}$ is such that $\mathcal{F} = \mathcal{I} + \mathcal{O}$. From a practical point of view, the $\mathcal{I}$ of each voxel corresponds to the number of times that the projection test is passed in a silhouette ($S(c)$) but not passed in the projection of the VH ($VH_{proj}(c)$): $PT_c(voxel, S(c)) = true \neq PT_c(voxel, VH_{proj}(c))$, being $voxel \in$ IV. For instance, voxels corresponding to object 1 in Figure 7(c), have $\mathcal{I} = 1$, for being in the inconsistent cone *camA→obj1*.

  In the IV, the number of inconsistencies ($\mathcal{I}$) is bounded by:

$$1 \leq \mathcal{I} \leq C - \mathcal{O} - 1, \tag{7}$$

  where the lower bound is due to the fact that all the voxels of the IV have been intersected with at least one inconsistent cone; and where the upper bound has to be lower or equal to $C - 1$, since C inconsistencies would correspond to a foreground detection of the Visual Hull. Moreover, the number of occlusions ($\mathcal{O}$) also has to be subtracted ($C - \mathcal{O} - 1$), since occlusions are only produced when voxels are intersected with consistent visual cones.
- Finally, a voxel can also be associated with the number of views where it projects to background ($\mathcal{B}$). Note that $\mathcal{B} = C - \mathcal{F}$, and therefore $\mathcal{B} = C - \mathcal{I} - \mathcal{O}$. The number of background projections ($\mathcal{B}$) can be computed by counting the number of silhouettes ($S(c)$) where $PT_c(voxel, S(c)) = false$, being $voxel \in$ IV. For instance, in Figure 7(c), $\mathcal{B} = 1$, for being in the inexistent cone *camC→obj1*.

The bounds on the number of background detections ($\mathcal{B}$) are: $1 \leq \mathcal{B} \leq C - \mathcal{O} - 1$. Besides using a similar reasoning as with the number of inconsistencies ($\mathcal{I}$), the expression can also be simply deduced by using inequality (7) substituting $\mathcal{I}$ with $\mathcal{I} = C - \mathcal{B} - \mathcal{O}$.

Some further considerations regarding $\mathcal{F}$, $\mathcal{I}$, $\mathcal{O}$ and $\mathcal{B}$ can be derived: Interestingly, the number of inconsistent projections ($\mathcal{I}$) in a voxel are due to either having had false alarms in $\mathcal{I}$ silhouettes or due to having had misses in $\mathcal{B}$ silhouettes, where $\mathcal{B} = C - \mathcal{I} - \mathcal{O}$.

As $\mathcal{I}$ rises, $\mathcal{B}$ falls, and therefore the probability of having $\mathcal{B}$ simultaneous misses is increased while the probability of having $\mathcal{I}$ simultaneous false alarms is decreased (see Figure 3(a) & Figure 3(c), respectively). Based on this reasoning, optimal threshold $T^\star$ has to be such that if $\mathcal{I} \geq T^\star$, the voxel is better explained as Shape (with $C - \mathcal{I} - \mathcal{O}$ misses) than Background (with $\mathcal{I}$ false alarms):

$$\begin{aligned} &\mathcal{I} \geq T^\star \Rightarrow \text{decide } \textit{Shape} \\ &\mathcal{I} < T^\star \Rightarrow \text{decide } \textit{Background} \end{aligned} \tag{8}$$

In order to find $T^\star$, first we have to express which is the probability of voxel miss-classification for any $P(Err_{3D}[T])$ so that $T^\star$ is that one which minimizes it:

$$T^\star = \operatorname*{argmin}_T P(Err_{3D}[T]) \tag{9}$$

Similarly as with the voxels in the VH, a voxel in the IV is miss-classified if it is wrongly classified as Shape (false alarm) or if it is wrongly classified as Background (miss), as expressed in (1).

Let's first examine the probability of false alarm ($P(FA_{3D})$). A false alarm in a voxel happens when a voxel is classified as part of the Shape, while in fact it forms part of the Background. If the voxel forms part of the Background, then all inconsistencies correspond to false alarms of the projection test. Since shape classification occurs when $\mathcal{I} \geq T$:

$$P(FA_{3D}) = \sum_{i=max(T,1)}^{C-\mathcal{O}-1} \binom{C}{i} P(FA_{2D})^i (1 - P(FA_{2D}))^{C-i}, \tag{10}$$

corresponding to the summation of all possible combinations that trigger a false alarm in a voxel, and assuming equiprobable $P_i(FA_{2D}) = P(FA_{2D})$ in all views (i).

Note that the combinations are bounded by the upper ($C - \mathcal{O} - 1$) and lower (1) bounds on the number of possible inconsistencies (see inequality 7), confirming previous considerations regarding the influence of occlusions. Also note that the expression is correctly expressed independently of the chosen T, even if the chosen value is out of the interval where the number of inconsistencies are possible.

The opposite miss-classification case in the IV is having a miss in a voxel. This is the case when a voxel is classified as part of the Background, while in fact it forms part of the Shape. Since a voxel is wrongly classified as Background if $\mathcal{I} < T$, and since $\mathcal{I} < T \iff \mathcal{B} \geq C - \mathcal{O} - T + 1$, then the probability of miss $P(M_{3D})$ in the IV can be expressed in a similar manner as with false alarms:

$$P(M_{3D}) = \sum_{i=max(C-\mathcal{O}-T+1,1)}^{C-\mathcal{O}-1} \binom{C}{i} P(M_{2D})^i (1 - P(M_{2D}))^{C-i}, \tag{11}$$

where $P(M_{2D})$ corresponds to the probability that the projection test has not been passed by error, and assuming equiprobable $P_i(M_{2D}) = P(M_{2D})$ in all views (i).

Once the probability of voxel miss-classification has been expressed, $T^\star$ can be easily obtained by doing an exhaustive search of the minimum $P(Err_{3D})$ over all possible T for each case of occlusion as shown in algorithm 3.

for all Cases of Occlusion: $o = 0 \cdots C - 1$ **do**
    $MinPerr \leftarrow 1$
    for all Possible Number of Inconsistencies: $i = 1 \cdots C - o - 1$ **do**
        if $P(Err_{3D}\,[T = i, \mathcal{O} = o)]) \leq MinPerr)$ **then**
            $T^\star[o] \leftarrow i$
            $MinPerr \leftarrow P(Err_{3D}\,[T = i, \mathcal{O} = o)])$

Algorithm 3: Optimal thresholds for all cases of occlusion: $T^\star[o]$. Note that $T^\star[o]$ will take different values for each voxel depending upon whether $P(M_{2D})$ or $P(FA_{2D})$ also depend on the voxel.

Note that the decisions on pixels which are not in the Visual Hull are made considering as foreground only the original Visual Hull. However, relabeling of a voxel in the inconsistent Hull could affect the decisions taken on other voxels (for instance in counting occlusions). The optimal solution should be made by considering all the possible interactions of all voxels, which would be an intractable problem.

Require: Silhouettes: $S(c)$, $T^{\star}[o]$, VH, a Proj. Test Function: $PT_c(voxel, Silhouette)$

```
Project the VH to all the camera views: VH_proj(c)
for all voxel do
  i ← 0
  o ← 0
  for all c do
    if PT_c(voxel, S(c)) is true then
      if PT_c(voxel, VH_proj(c)) is false then
        i ← i + 1
      else
        o ← o + 1
  if i > 0 then
    if i ≥ T*[o] then
      UH(voxel) ← 3D Foreground
    else
      UH(voxel) ← 3D Background
```

Algorithm 4: SfIS algorithm

4 Real-Time SfIS

SfIS can be very fast, once the optimal thresholds have been computed for each possible case of occlusion and stored in a lookup table (LUT). Real-time operation of SfIS can be achieved when using it in combination with fast projection tests. Often, the One Pixel Projection Test is used for being fast and simple. However, LUTs cannot be used when probabilities of 2D miss and false alarm of the projection test change over time ($P(FA_{Pix}(t)$ and $P(M_{Pix}(t))$. For example, when a mixture of Gaussians is used to model the Background, the probabilities of miss and false alarm of the pixels depend on the variances of the Gaussians, which are constantly changing over time.

Under these circumstances, it is important to have a fast search strategy that can compute the optimal thresholds on-line.

4.1 A Fast Threshold Search Approach

The method presented here is focused on a fast implementation of SfIS using the One Pixel Projection Test. However, we develop the equations for the more general case of any projection test which is equiprobable with respect to all views.

First off, since $P(Err_{3D}[T])$ is not continuous, it cannot be minimized by differentiating it. However, fast search of $T^\star$ can be achieved if the problem can be constrained into finding the minimum of $P(Err_{3D}[T])$ in a strictly convex interval L. In other words, if we can guarantee that $P(Err_{3D}[T])$ is strictly convex in L under certain conditions, and provided that these conditions are reasonable, then there will always exist a global minimum in L, which will be fast to obtain.

Following, we propose some sufficient conditions which guarantee the strict convexity of $P(Err_{3D}[T])$, with respect to $T \in \mathbb{Z}$ in the interval of interest L: first, we find which is the interval, and then, we obtain the conditions.

In order to find the L of interest, it is important to remember that the range of possible inconsistencies in a voxel in the IV is $\mathcal{I} \in [1, C - \mathcal{O}]$. This is the reason why, in the IV, $P(Err_{3D}[T])$ has constant values for $T \leq 1$ and $T \geq C - \mathcal{O}$, corresponding to the probabilities of always deciding Shape or always deciding Background, respectively. Since strict convexity of a function can only occur in the interval where the function is not constant, the interval of convexity of $P(Err_{3D}[T])$ has to be: $L \in]1, C - \mathcal{O}[$.

Once that L has been determined, we only have to seek the conditions that make $P(Err_{3D}[T])$ strictly convex in the interval. In general, a function $f[x]$ in $\mathbb{Z}$ is strictly convex [39,40] if it can be expressed as in inequality (12).

$$f[x-1] + f[x+1] > 2f[x] \tag{12}$$

And since conditions of strict convexity have to be found assuming that the projection tests are equiprobable in all camera views, the working expression of $P(Err_{3D}[T])$ is:

$$\begin{aligned} P(Err_{3D}[T]) = \; & P_S \underbrace{\sum_{i=max(C-\mathcal{O}-T+1,1)}^{C-\mathcal{O}-1} \binom{C}{i} P(M_{2D})^i (1-P(M_{2D}))^{C-i}}_{P(M_{3D})} + \\ & (1-P_S) \underbrace{\sum_{i=max(T,1)}^{C-\mathcal{O}-1} \binom{C}{i} P(FA_{2D})^i (1-P(FA_{2D}))^{C-i}}_{P(FA_{3D})} \end{aligned} \tag{13}$$

Then, strict convexity of $P(Err_{3D}[T])$ occurs if (13) satisfies (12):

$$P(Err_{3D}[T-1]) + P(Err_{3D}[T+1]) > P_S P(M_{3D}) + (1-P_S)P(FA_{3D}).$$

In the interval $L \in]1, C - \mho[$, the inequality above can be expressed as:

$$\underbrace{P_S \left(\binom{C}{T-1} P(FA_{2D})^{T-1} (1 - P(FA_{2D}))^{C-T+1} - \binom{C}{C-\mho-T+1} P(M_{2D})^{C-\mho-T+1} (1 - P(M_{2D}))^{\mho+T-1} \right)}_{T_1} +$$
$$\underbrace{(1 - P_S) \left(\binom{C}{C-\mho-T} P(M_{2D})^{C-\mho-T} (1 - P(M_{2D}))^{\mho+T} - \binom{C}{T} P(FA_{2D})^{T} (1 - P(FA_{2D}))^{C-T} \right)}_{T_2} > 0,$$

where $P(M_{3D})$ and $P(FA_{3D})$ are canceled.

As a matter of fact, since we only need to find a *sufficient* condition of strict convexity of $P(Err_{3D}[T])$, we can separate the left term of the inequality into two terms (T_1, T_2), and seek the conditions that make both terms larger than 0. Forcing the first term to be greater than 0, can be expressed as follows:

$$T < \frac{(C - \mho + 1) P(M_{2D})^{-1} (1 - P(M_{2D})) - \mho}{1 + P(M_{2D})^{-1} (1 - P(M_{2D}))} \tag{14}$$

In order to guarantee that inequality (14) is satisfied in L, we can impose a stricter condition on T, by replacing it with $C - \mho$ which is larger than the largest possible value that T can take in the interval L:

$$C - \mho < \frac{(C - \mho + 1) P(M_{2D})^{-1} (1 - P(M_{2D})) - \mho}{1 + P(M_{2D})^{-1} (1 - P(M_{2D}))} \Leftrightarrow C < \frac{1}{P(M_{2D})} - 1, \tag{15}$$

which is stricter than condition (14).

Note that if condition (15) is satisfied, then condition (14) is also satisfied, and therefore T_1 is greater than 0 in L.

We can do a similar reasoning with T_2, which is larger than 0 if:

$$C < \frac{1}{P(FA_{2D})} - 1 \tag{16}$$

Finally, $P(Err_{3D}[T])$ can be said to be strictly convex in $L \in]1, C - \mho[$ if conditions (15) and (16) hold together, which can be expressed in a single inequality as follows:

$$C < \frac{1}{max(P(FA_{2D}), P(M_{2D}))} - 1, \tag{17}$$

which is usually satisfied in all typical scenarios. For instance, $P(Err_{3D}[T])$ is strictly convex with respect to T, when less than 9 cameras are used even if miss-classification probabilities are high ($P(FA_{2D}) = P(M_{2D}) = 0.1$).

Note that if condition (17) is not satisfied, then we cannot guarantee whether $P(Err_{3D}[T])$ is convex or not, and algorithm 3 has to be used. However, when the condition holds, which is often the case, we can make use of some of the properties of strictly convex functions:

If $P(Err_{3D}[T])$ can be said to be strictly convex, then its central difference $\delta P(Err_{3D}[T])$ will correspond to a sequence of sorted elements in increasing order. And the element in the sorted sequence which is closest to zero, will correspond to the minimum of $P(Err_{3D}[T])$:

$$\delta P(Err_{3D}[T]) = P(Err_{3D}[T+1] - P(Err_{3D}[T-1]) \tag{18}$$

where δ is the central difference operator.

Observe that if the sequence $\delta P(Err_{3D}[T])$ is sorted, we can approach the closest value of 0 from the left side, by checking whether the midpoint of the sequence is larger than 0, eliminating half of the sequence from further consideration. The binary search [41] is an algorithm that repeats this procedure, halving the size of the remaining portion of the sequence each time. The complexity of the search operation in the binary search is $O\log_2(n)$, because at each test one half of the search space is discarded. Furthermore, $\delta P(Err_{3D}[T])$ can be computed $\frac{4}{C-\mathcal{O}-1}$ times faster than $P(Err_{3D}[T])$, since the sum over all possible cases of 3D false alarm and miss ($\sum^{C-\mathcal{O}-1}$) does not have to be computed.

In conclusion, the method described achieves the optimal solution $T^\star$ in $O\log_2(\frac{4n}{C-\mathcal{O}-1})$ time, which is faster than the linear search approach described in algorithm 3, which only achieves the solution in $O(n)$ time. It is noteworthy that the method proposed improves drastically as the size of the array, i.e. the number of cameras, is increased.

Finally, refer to algorithm 5 for the detailed implementation of the method, considering a left approach to the optimal solution $T^\star$. Note that the optimal solution has to be found for every possible case of occlusion that may occur in the scene.

5 Results

In order to fully evaluate SfIS, two types of results are presented. First, we present the theoretical improvements of SfIS over SfS in the IV. Second, we show some $VH \cup IV$ reconstructions and projections using synthetic and real data.

Require: $\delta P(Err_{3D}[T,o])$
for all Cases of Occlusion: $o = 0 \cdots C-1$ **do**
  $left \leftarrow 1$
  $right \leftarrow C - \mathcal{O}$
  while $left \leq right$ **do**
    $index \leftarrow \left\lceil \frac{left+right}{2} \right\rceil$
    if $\delta P(Err_{3D}[T = index, \mathcal{O} = o]) > 0$ **then**
      $right \leftarrow index - 1$
    else
      $left \leftarrow index + 1$
  return $T^{\star}[o] = \mathrm{argmin}_{T=\{index,index+1\}}\, \delta P(Err_{3D}[T, \mathcal{O} = o])$

Algorithm 5: Binary search of $T^{\star}[o]$ for all cases of occlusion (o).

5.1 Theoretical Improvements

It is important to keep in mind that SfIS is focused on minimizing the probability of Shape miss-classification in the IV in terms of *consistency*. This is the reason why all points which belong to consistent zones are considered to be error-free.

So, in order to compare the errors of SfS and SfIS as fairly as possible, let us first rewrite the expressions of error of SfS in the IV, assuming that there cannot be consistent miss-classifications. The new error rate is lower than the one presented in section 2, which considered that miss-classifications could also be consistent. However the reformulation is necessary in order to not unfairly worsen the results of SfS in front of SfIS:

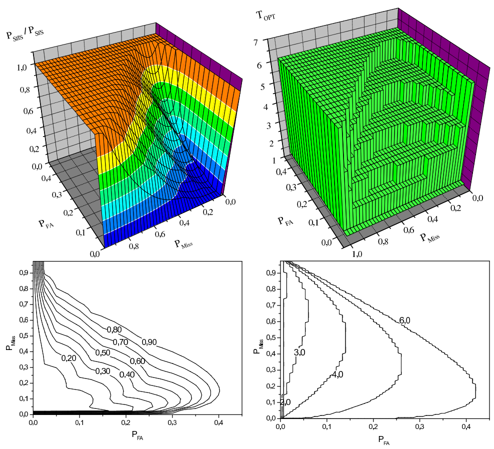
$$\begin{aligned}
&P_{SfS}(Err_{3D}) = P_B P_{SfS}(FA_{3D}) + P_S P_{SfS}(M_{3D})\\
&P_{SfS}(FA_{3D}) = 0\\
&P_{SfS}(M_{3D}) = \sum_{i=1}^{C-\mathcal{O}-1} \binom{C}{i} P(M_{2D})^i (1 - P(M_{2D}))^{C-i},
\end{aligned} \tag{19}$$

where $P(M_{2D})$ corresponds to the probability that the projection test has not been passed by error, and assuming equiprobable $P_i(M_{2D}) = P(M_{2D})$ in all views (i). Note that the upper bound is $C - \mathcal{O} - 1$, corresponding to the maximum number of background detections possible in the IV.

Figure 8(a) shows the $\frac{P_{SfIS}(Err_{3D})}{P_{SfS}(Err_{3D})}$ probability ratio, assuming that an equiprobable projection test is being used and that there have not been occlusions. Note that the ratio is always below 1, meaning that the probability of voxel miss-classification in SfIS is always lower than in SfS.

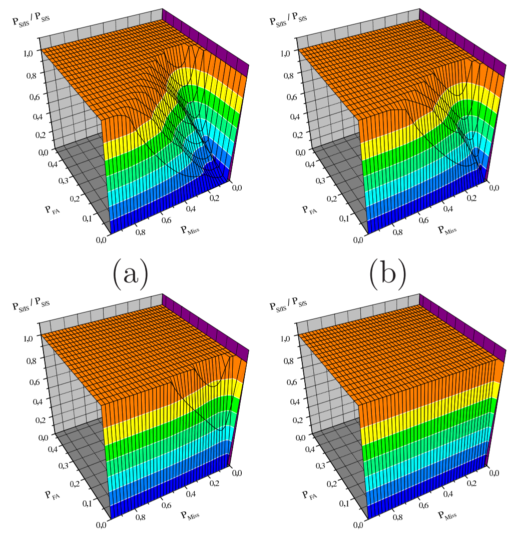
An aspect of interest of SfIS is that it behaves as traditional SfS when (1) $P(FA_{2D})$ is high or (2) $P(M_{2D})$ is low: In the first case, when there are high chances of 2D false alarm, SfIS mimics SfS in order to not incorporate 3D false alarms. In the second case, when $P(M_{2D})$ is very low, SfIS does not interfere either, since SfS is the best reconstruction method when there are not misses. In both cases, $T^{\star} = C - \mathcal{O}$ (see 8(b)), forcing the system to always decide Background, and leaving an empty IV.

Figure 9 shows how $\frac{P_{SfIS}(Err_{3D})}{P_{SfS}(Err_{3D})}$ varies with different number of occlusions. Note that as $\mathcal{O}$ rises, SfIS has less room for maneuver. In any case, even when $\mathcal{O} = C - 1$, the probability of miss-classification of SfIS is never worse than with SfS.



(a) (b)

Figure 8. The ratio $\frac{P_{SfIS}(Err_{3D})}{P_{SfS}(Err_{3D})}$, and $T^{\star}$ for different values of $P(FA_{2D})$ and $P(M_{2D})$. Results are shown considering a set-up of 6 cameras with $P_B = 0.9, P_S = 0.1$. In this case, it is assumed that there have not been occlusions ($\mathcal{O} = 0$). Note that $T^{\star} = C$ when there are not misses or when the probability of 2D false alarm is high. If $T^{\star} = C$ then $P_{SfS}(Err_{3D})$ is equivalent to $P_{SfIS}(Err_{3D})$.



(a) (b) (c) (d)

Figure 9. $\frac{P_{SfIS}(Err_{3D})}{P_{SfS}(Err_{3D})}$ probability ratio when there have been 1, 2, 3, and 4 occlusions, using the same set-up as in Figure 8.

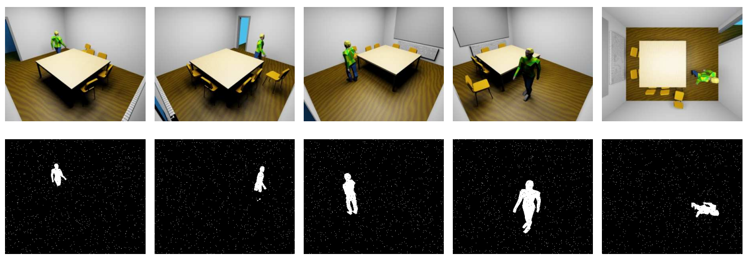
5.2 *Empirical Results*

Following we present two different experiments showing the performance of SfIS in front of SfS. The first experiment consists in the reconstruction of the VH and $VH \cup UH$ from a set of synthetic images using different projection tests. The experiment includes quantitative results of the algorithms. In the second experiment, we use real-word images obtained in the smart-room of our lab to show results which can be straight away evaluated from simple observation.

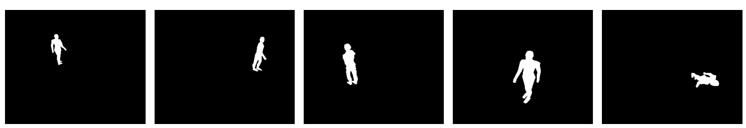
5.2.1 Results with Synthetic Images

In order to obtain quantitative results of the algorithm, we have employed a set of synthetic images because we can arbitrarily add/remove random noise and occlusions to them and results can be objectively evaluated using the ground truth.

The set of images which we have used are shown in Figure 10. The first row of images depicts the synthesized scene used in the experiments. In the pictures, a table partially occludes the bottom part of a person in the first two views. The second row of images shows the corresponding set of silhouettes. Note that the silhouettes have some misses and false alarms which have been artificially added. Finally, in the row at the bottom, the silhouettes corresponding to the noise-free and occlusion-free consistent scene are shown.



Original images and silhouettes with artificial noise ($P(M_{2D}) = P(FA_{2D}) = 0.01$).



Original synthetic silhouettes without occlusions or artificial noise.

Figure 10. Set of synthetic images and silhouettes. The dataset, which is a courtesy of J.C. Pujol from the Carlos III University of Madrid, consists of 5 sequences of frames of 352x288 pixels. In the scene, the cameras, table and chairs are positioned resembling the set-up of the smart-room of the UPC.

The evaluation process is performed as follows. First, a reconstruction from the set of consistent silhouettes, corresponding to the third row in Figure 10, is obtained to be used as the Ground Truth ($G.T.$). Then, SfS and SfIS algorithms are employed to reconstruct 3D Shapes using the bogus silhouettes of the second row in Figure 10. Finally, these Shapes are confronted with that one which was reconstructed using the consistent set of silhouettes.

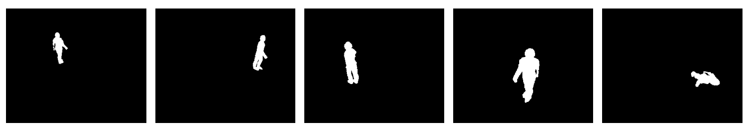
In order to evaluate the performance of the system, we have employed the verification measures that are commonly used in the information retrieval field:

$$\text{Recall} = \frac{\#\text{correct Shape detections}}{\#\text{correct Shape detections} + \#\text{misses}}. \tag{20}$$

$$\text{Precision} = \frac{\#\text{correct Shape detections}}{\#\text{correct Shape detections} + \#\text{false Shape detections}}. \tag{21}$$

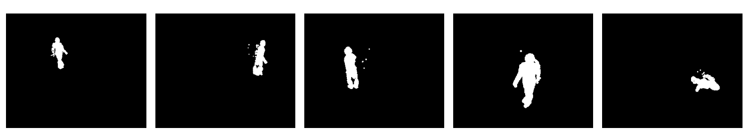
$$\text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \tag{22}$$

Recall measures how well the classifier detects voxels that form part of the Shape and precision measures how well it weeds out the voxels in the background. A well balanced system should have high, similar values of both recall and precision.



VH projection using SfS with the One Pixel Projection Test using silhouettes without occlusions or noise.

VH projection using SfS with the One Pixel Projection Test using the bogus silhouettes.



$VH \cup UH$ projection using SfIS with the One Pixel Test using the bogus silhouettes.

Figure 11. SfIS vs. SfS using 5 cameras, and the One Pixel Projection Test.

Figure 11 shows three different types of reconstructions using the One Pixel Projection Test. (1) In the first row, the projections of the Shape reconstructed from the noise-free and occlusion-free images are observed. In this case, the standard SfS algorithm has been used to obtain the voxelized scene. Since the Shape has been obtained from the set of consistent silhouettes, the labeled voxels are used as the Ground Truth ($G.T.$) for comparison in Table 2. (2) The second row of images corresponds to results of SfS using the noisy and partially occluded silhouettes. Note that most of the errors correspond to 3D misses, as shown in Table 2. (3) The row at the bottom shows the projection of the Shape using the proposed SfIS method. Note that SfIS is able to recover that part of the Shape that the table occluded in the first two views, and also improves the detection in case of global noise. Besides, observe that SfIS is also successful in not incorporating more false alarms than the recovered misses, as shown in Table 2. Also note that some of the wrongly reconstructed voxels can be easily removed in a further stage using simple morphological operations such as opening by reconstruction. A 3D area opening of size 4 has been used to obtain the results of the third column in Table 2.

Table 2 offers another interesting result. Note that SfIS is not only better than SfS w.r.t. the F-measure, but it also has similar precision and error rates implying an unbiased treatment of error types in 3D.

Table 2
Results using the One Pixel Projection Test

	Ground truth	SfS	SfIS	SfIS-filt
# Correct detections	5381	4371	**5071**	**5071**
# False alarms	0	**2**	391	330
# Misses	0	1010	**310**	**310**
Recall	1	0.81	**0.94**	**0.94**
Precision	1	**0.99**	0.93	0.94
F-measure : $\frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}}$	1	0.90	0.93	**0.94**

We want to remark that the data on Table 2 is only provided to validate the implemented system. That is, we confirm, that, as imposed in the design of the system, we reduce the total error and we can balance the two kind of errors (false alarms and misses). We could consider SfS as a version of SfIS where the threshold taken for the voxels of the Inconsistent Volume was set to a fixed number ($C^\star$, the number of cameras). However, in most applications, the best threshold is the one than minimizes the total error, that is the one proposed in last section.

We have experimented other projection tests, obtaining similar results. Independently of the projection test, SfIS is always useful because it balances errors between false alarms and misses and produces the lowest possible total error.

5.2.2 Results with Real-World Images

In Figure 12, a real world scenario is shown. In this case, the foreground segmentation has been done using [6], and we have added some additional false alarms in (a) and (d).

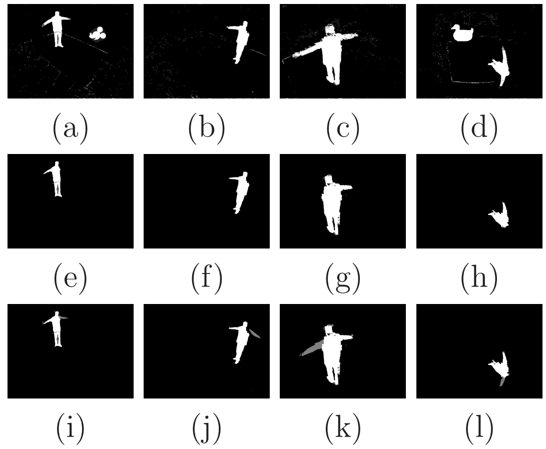


Figure 12. Silhouettes, projection of the VH and projection of the $VH \cup UH$, in first, second and third row, respectively.

The experiment has been performed using a very high-resolution volume, employing voxels with edge size of less than 5 mm. The underlying idea is to guarantee that the projection of the splat of any voxel in the shape is comprised within a pixel in all the silhouettes. Therefore, the reconstruction is independent of the projection test and $P(FA_{2D})$ and $P(M_{2D})$ concur with the probabilities of FA and Miss of the background learning technique. In this case, we are assuming $P_{FA}(2D) = P_M(2D) = 0.1$, and P_B has been selected based on the percentage of voxel occupancy in the VH.

In (b), the silhouette's left arm has not been detected due to the similar color to its background counterpart. The second row of images shows the projection of the VH, reconstructed using the standard voxel-based SfS algorithm. Note that the miss-detection in (b) has been propagated to the rest of silhouettes. The row of images on the bottom shows the projection of the $VH \cup UH$ in white and gray, respectively. Observe that the projection of the arm is recovered, even in (j), while remaining unaffected to the artificial false alarms.

The experiment shows that SfIS can be used to recover some of the errors produced in the 2D foreground segmentation techniques by exploiting the redundancy present in a multi-camera setup. On the contrary, SfS does not only fail to recover these types of errors but it actually worsens all the silhouettes by propagating the 2D misses from one view to the rest of views.

To complete the experiments, we have have considered it appropriate to include here a set of tests comparing SfIS with other approaches using real world video sequences. This time, the experiment has been performed using a low-resolution volume, employing voxels with edge size 2.5 cm, therefore prioritizing fast 3D detections over a more accurate Shape.

Also, since we are using real-world images with imprecise calibration, we have opted to indirectly evaluate the performance of the reconstruction methods.

To do so, we have compared the projection of the 3D volumes with a set of five manually classified silhouettes from images that have been randomly selected in the video sequence. These manually labeled silhouettes are the ground truth in this case.

Three different techniques have been compared. The first technique is a version of SfS where a voxel is not classified as Shape if there are *more than one* views where its Projection Test fails. We identify this method as *SfS* $C - 1$ *intersections* in our experiments. The second evaluated technique is traditional SfS. The third tested method is SfIS. In this experiment, we have employed the One Pixel Projection Test in all the methods for a fair comparison.

The pixel models employed for 2D classifications are a single Gaussian per pixel for the background and a uniform distribution for the foreground. 2D classifications are obtained using MAP and the models are updated using EM.

For visual inspection purposes, we present two figures (Figure 13 and Figure 14) with results corresponding to different times and camera views of a scene. The original image, the 2D segmentation and the projections of the Shapes obtained with the methods under evaluation are shown.

In Figure 13, the images corresponding to camera 2 in frame number 175 are shown. Note that the 2D only segmentation (2nd column, 1st row) -not using 3D redundancy information- has failed due to the similar colors of the person in the foreground and the clutter in the background.

Similar problems are observable in Figure 14. The figure corresponds to frame 650 and shows two out of the five camera views used in all the methods. Note that 2D misses in a view are transferred to the rest of views in the SfS approach. The SfS $C - 1$ approach does not propagate 2D misses but incorporates many false alarms conducing to larger Shapes and silhouettes' projections. As it can be observed from the images, SfIS is a good approach for not propagating 2D misses as well as for not incorporating many false alarms.

View of Camera 2	2D only Segmentation	SfS $C - 1$ intersections	SfS	SfIS

Figure 13. Silhouettes and 3D volumetric projections corresponding to frame 175 with different techniques using the One Pixel Projection Test.

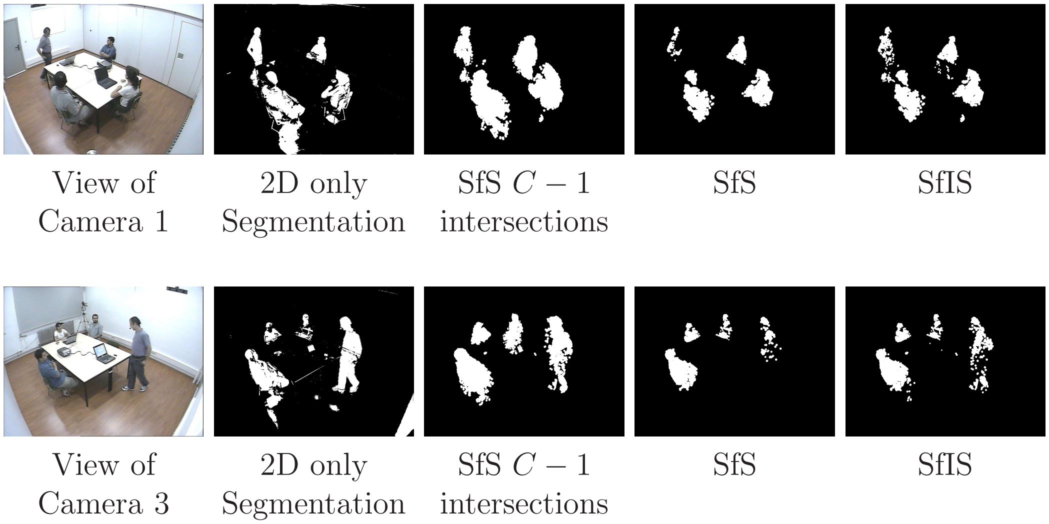


Figure 14. Silhouettes and 3D volumetric projections corresponding to frame 650 with different techniques using the One Pixel Projection Test.

Quantitative results of this experiment are presented in Table 3. These results have been obtained by averaging the number of 2D false alarms, 2D correct detections and 2D misses over a set of projected reconstructions. These projections correspond to the five views where the silhouettes were manually labeled to be the ground truth, as previously commented.

Table 3
System Evaluation through the Projection of 3D Reconstructions in Video Sequences

	Ground truth	SfS ($C-1$ int.)	SfS	SfIS
# Correct foreground det.	32270	**27471**	15023	20445
# False alarms	0	29760	**5077**	7529
# Misses	0	**4808**	13256	11834
Recall	1	**0.85**	0.53	0.63
Precision	1	0.48	**0.75**	0.73
F-measure : $\frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}}$	1	0.62	0.62	**0.68**

Some interesting conclusions can be extracted from the table. Note that the SfS $C-1$ approach has a highest recall rate. Indeed, it also has a very large number of false alarms and, therefore, a poor precision rate, but it is a good method if we want to be sure to detect the foreground voxels when they exist.

In contrast, traditional SfS is very precise, even more than SfIS. SfS detects fewer voxels but it is very good at asserting that those voxels form part of the Shape.

SfIS is the most balanced method. It has high precision and recall rates and its F-measures is the best. In conclusion, SfIS has the best F-score of them, as simple visual inspection of the images confirms and it is the method that performs best when operating with video sequences [4] .

6 Conclusion

In this paper we have presented a novel scheme for effective Shape from Silhouette using sets of inconsistent silhouettes as usually found in practical scenarios. The scheme exploits the consistency principle, and performs an error detection and correction procedure of the most probable consistent silhouettes according to the available data.

First, we have introduced a method to determine the IV, i.e., the volumetric zones leading to inconsistent regions in the silhouettes. Then, we have described a voxel-based technique -which works with any projection test- to enumerate the number of inconsistent cone intersections. Finally, we have proposed a method to obtain the minimum number of inconsistent cone intersections $T^\star$ that have to be produced so that it can be determined that an object was wrongly missed by a SfS technique. Threshold $T^\star$ is taken after minimization of the probability of voxel miss-classification. In addition, we have stated the conditions in which a very fast implementation of SfIS is possible.

SfIS has proved to be an effective 3D reconstruction tool. We have given theoretical prove that SfIS miss-classification probability is lower than the one using SfS. Experimental results have also been carried out showing that SfIS does not only reduce the number of errors but it is also successful in balancing errors between 3D false alarms and misses, contrary to conventional SfS that mainly introduces only 3D misses. Indeed, SfIS introduces false alarms to the Shape. However false alarms are introduced only to the extent that global error is lower than without them. SfIS has been show to be effective for reducing both Gaussian nature and deterministic errors.

SfIS can be used in at least two different manners. (1) To extract better volumes by minimizing the effects that inconsistencies have over the reconstructed Shape. (2) To recover errors in the silhouettes from informed decisions made at the volume level, where individual detections at the 2D level are compared for consistency. Contrarily, conventional SfS does not only fail to recover er-

[4] Due to the great number of images resulting from all the methods compared, the number of camera views and the large time interval evaluated, it is not possible to show here the complete set of resulting images. However, the video sequences with all the evaluated methods at all the cameras views can be obtained in `http://gps-tsc.upc.es/imatge/_jl/`

rors in the silhouettes but it worsens the silhouettes by propagating 2D misses from one view to the others.

Some of the future possibilities of SfIS include giving feedback to the background learning techniques to make more trustworthy background models. Also,the combination of SfIS with methods that use the silhouettes' probabilities instead of using binary silhouettes for the first voxel occupancy decision can further improve the results. Although these methods [28,35,30] provide a voxel occupancy decision which takes into account the multi-camera information, inconsistencies still remain due for instance to occlusions in one of the views. Thus, the Visual Hull created by these methods can be further refined by classifying the Inconsistent Volume that they produce.

This work was partially developed within the framework of Spanish Project VISION CENIT 2007-1007.

References

[1] Bruce G. Baumgart. *Geometric Modeling for Computer Vision*. PhD thesis, CS Department, Stanford University, October 1974. AIM-249, STAN-CS-74-463.

[2] A. Laurentini. The Visual Hull: A new tool for contour-based image understanding. In *Proceedings of Seventh Scandinavian Comperence on Image Processing*, pages 993–1002, 1991.

[3] A. Laurentini. The Visual Hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):150–162, 1994.

[4] Aldo Laurentini. How far 3D shapes can be understood from 2D silhouettes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(2):188–195, 1995.

[5] Andrea Bottino and Aldo Laurentini. Introducing a new problem: Shape from Silhouette when the relative positions of the viewpoints is unknown. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(11):1484–1493, 2003.

[6] Chris Stauffer and W. Eric L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[7] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Paul Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[8] T. Horpraset, D. Harwood, and L. Davis. A statistical approach for real-time robust background subtraction and shadow detection. In *Proceedings of International Conference on Computer Vision*. IEEE Computer Society, 1999.

[9] Stephen J. McKenna, Sumer Jabri, Zoran Duric, Azriel Rosenfeld, and Harry Wechsler. Tracking groups of people. *Computer Vision and Image Understanding*, 80(1):42–56, 2000.

[10] Haritaoglu, D. Harwood, and L. Davis. W4: Real time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, August 2000.

[11] Ahmed Elgammal, Ramani Duraiswami, David Harwood, and Larry S. Davis. Non-parametric model for background subtraction. In *Proceedings of International Conference on Computer Vision*. IEEE Computer Society, Sept 1999.

[12] Anurag Mittal and Larry S. Davis. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo. In *Proceedings of European Conference on Computer Vision*, pages 18–36, London, UK, 2002. Springer-Verlag.

[13] S. Khan and M. Shah. Tracking people in presence of occlusion. In *Proceedings of Asian Conference on Computer Vision*, 2000.

[14] Liyuan Li, Weimin Huang, Irene Y. H. Gu, and Qi Tian. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Transactions on Image Processing*, 13(11):1459–1472, 2004.

[15] Alvy Ray Smith and James F. Blinn. Blue screen matting. In *Proceedings of International Conference and Exhibition on Computer Graphics and Interactive Techniques*, pages 259–268, New York, NY, USA, 1996. ACM Press.

[16] Yung-Yu Chuang, Brian Curless, David H. Salesin, and Richard Szeliski. A Bayesian approach to digital matting. In *Proceedings of Computer Vision and Pattern Recognition*, volume 2, pages 264–271. IEEE Computer Society, December 2001.

[17] O. Javed and M. Shah. Tracking and object classification for automated surveillance. In *Proceedings of European Conference on Computer Vision*, pages 343–357, 2002.

[18] L-Q Xu, J. L. Landabaso, and B Lei. Segmentation and tracking of multiple moving objects for intelligent video analysis. *BT Technology Journal*, 22(3):140–150, July 2004.

[19] J. L. Landabaso, M. Pardàs, and L.-Q. Xu. Shadow removal with blob-based morphological reconstruction for error correction. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, PA, USA, March 2005. IEEE Computer Society.

[20] L. Wixson. Detecting salient motion by accumulating directionally-consistent flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):774–780, 2000.

[21] M. Piccardi. Background subtraction techniques: a review. In *Proceedings of IEEE SMC International Conference on Systems, Man and Cybernetics*, volume 4, pages 3099–3104, The Hague, The Netherlands, Oct 2004.

[22] Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J. Gortler, and Leonard McMillan. Image-based visual hulls. In *Proceedings of International Conference and Exhibition on Computer Graphics and Interactive Techniques*, pages 369–374, New York, NY, USA, 2000. ACM Press.

[23] Ari Rappoport and Steven Spitz. Interactive boolean operations for conceptual design of 3-d solids. In *Proceedings of International Conference and Exhibition on Computer Graphics and Interactive Techniques*, pages 269–278, New York, NY, USA, 1997. ACM Press.

[24] Jack Goldfeather, Jeff P M Hultquist, and Henry Fuchs. Fast constructive-solid geometry display in the pixel-powers graphics system. In *Proceedings of International Conference and Exhibition on Computer Graphics and Interactive Techniques*, pages 107–116, New York, NY, USA, 1986. ACM Press.

[25] Saied Moezzi, Arun Katkere, Don Y. Kuramura, and Ramesh Jain. Reality modeling and visualization from multiple video sequences. *IEEE Computer Graphics and Applications*, 16(6):58–63, 1996.

[26] Saied Moezzi, Li-Cheng Tai, and Philippe Gerard. Virtual view generation for 3D digital video. *IEEE MultiMedia*, 4(1):18–26, 1997.

[27] J. L. Landabaso and M. Pardàs. Foreground regions extraction and characterization towards real-time object tracking. In *Proceedings of Multimodal Interaction and Related Machine Learning Algorithms*, Lecture Notes in Computer Science. Springer, 2005.

[28] D. Snow, P. Viola, and R. Zabih. Exact voxel occupancy with graph cuts. In *Proceedings of Computer Vision and Pattern Recognition*, pages 345–353. IEEE Computer Society, 2000.

[29] Kong Man Cheung, Takeo Kanade, J.-Y. Bouguet, and M. Holler. A real time system for robust 3D voxel reconstruction of human motions. In *Proceedings of Computer Vision and Pattern Recognition*, volume 2, pages 714 – 720. IEEE Computer Society, June 2000.

[30] J. L. Landabaso and M. Pardàs. Cooperative background modelling using multiple cameras towards human detection in smart-rooms (invited paper). In *Proceedings of European Signal Processing Conference*, 2006.

[31] Michael Potmesil. Generating octree models of 3D objects from their silhouettes in a sequence of images. *Computer Vision, Graphics and Image Processing*, 40(1):1–29, 1987.

[32] Richard Szeliski. Rapid octree construction from image sequences. *Computer Vision, Graphics and Image Processing*, 58(1):23–32, 1993.

[33] Greg Slabaugh, Bruce Culbertson, Thomas Malzbender, and Ron Shafer. A survey of methods for volumetric scene reconstruction from photographs. In *International Workshop on Volume Graphics*, Stony Brook, New York, June 2001.

[34] C. R. Dyer. Volumetric scene reconstruction from multiple views. In *Foundations of Image Understanding*, pages 469–489. Kluwer, 2001.

[35] Jean-Sébastien Franco and Edmond Boyer. Fusion of multi-view silhouette cues using a space occupancy grid. In *Proceedings of International Conference on Computer Vision*. IEEE Computer Society, October 2005.

[36] K. Forbes, A. Voigt, and N. Bodika. Using silhouette consistency constraints to build 3D models. In *Proceedings of Fourteenth Annual South African Workshop on Pattern Recognition*. PRASA, 2003.

[37] Computers in the Human Interaction Loop (CHIL) EU Project.

[38] J. L. Landabaso, M. Pardàs, and J.R. Casas. Reconstruction of 3D shapes considering inconsistent 2D silhouettes. In *Proceedings of International Conference on Image Processing*. IEEE Computer Society, 2006.

[39] Kazuo Murota. Discrete convex analysis — exposition on conjugacy and duality. *Graph Theory and Combinatorial Biology. Bolyai Mathematical Society*, 7:253–278, 1999.

[40] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[41] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Second Edition*. The MIT Press, September 2001.