RECONSTRUCTION OF 3D SHAPES CONSIDERING INCONSISTENT 2D SILHOUETTES

J.L. Landabaso, M.Pardàs, J.R.Casas

Image Processing Group, Technical University of Catalonia, Barcelona, Spain

ABSTRACT

The Visual Hull is defined as the intersection of the visual cones formed by the back-projection of C 2D silhouettes into the 3D space. The set of 2D silhouettes is consistent if there exists at least one volume which exactly explains them.

Shape from Silhouette (SfS) is the general term used to refer to the techniques employed to obtain a volume from silhouettes, which are considered to be consistent. In this paper we extend the idea of SfS to be used with sets of inconsistent silhouettes resulting from inaccurate calibration and erroneous 2D silhouette extraction techniques. The method presented detects and corrects errors in the silhouettes based on the consistency principle, implying an unbiased treatment of false alarms and misses in 2D.

Index Terms— Shape from Silhouette, SfS, SfIS, Visual Hull, Inconsistent Hull, Reconstruction, 3D, Computer Vision, Smart Room

1. INTRODUCTION

Shape extraction from a set of silhouettes (binary masks of the objects 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. 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 silhouettes, and the VH is the maximal volume among the possible ones. Therefore, it follows that the VH exists if and only if the silhouettes are consistent. However, 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[3], chroma key techniques, etc. In spite of that, SfS methods have been designed in the past assuming that the silhouettes are consistent, reconstructing then only that 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 and adjusts the regions which are not, implying an unbiased treatment of all sorts of 2D errors, i.e., misses and false alarms. By contrast, other SfS systems usually treat differently the 2D errors on the basis of their type. In the following, we review which are the different types of 2D errors and how they affect the reconstructed shape.

This material is based upon work partially supported by the IST programme of the EU through the IP CHIL and Noe SIMILAR.

2. NOISE PROPAGATION TO THE 3RD DIMENSION

Silhouettes may contain false alarms or misses, corresponding to erroneous foreground detections or erroneous background detections, respectively.

In SfS, a false alarm in a view does not conduce to a false alarm in 3D unless the erroneous visual cone intersects simultaneously with other $C-1$ visual cones, where C is the total number of cameras. If the intersection is produced, then the shape is wrongly reconstructed, letting a consistent reconstruction with undetectable 2D false alarms. However, the shape is not reconstructed when at least one of the visual cones does not intersect. Then, there is at least one 2D false alarm which is detectable as it is inconsistent with the rest of silhouettes (see Fig. 2(b)).

Contrarily, a miss in 2D ineluctably conduces to a miss in 3D, meaning that the shape is not reconstructed no matter whether the shape is consistent or not (see Fig. 2(a)).

In conclusion SfS algorithms 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 sorted out based on a different principle; one that takes decisions based on 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.

2.1. Dealing with the 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 essentially three sorts of approaches.

The first general approach involves using voxel-based reconstructions to reduce the probability of voxel miss-classification. In [4], Cheung et al. propose an algorithm called SPOT which determines the minimum number of foreground pixels (Z_ϵ) which have to be detected inside the projection of a voxel to consider that the Projection Test is passed in a certain silhouette. 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*. Even though SPOT is an important step forward, it does not succeed in detecting deterministic errors, often consisting in large regions missed in a view when foreground objects have similar colors and texture to their background counterparts.

The second general approach[5, 6], 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 either if the silhouettes are consistent or not.

The last general approach uses multi camera information in terms of consistency constraints, providing tools for detecting determinis-

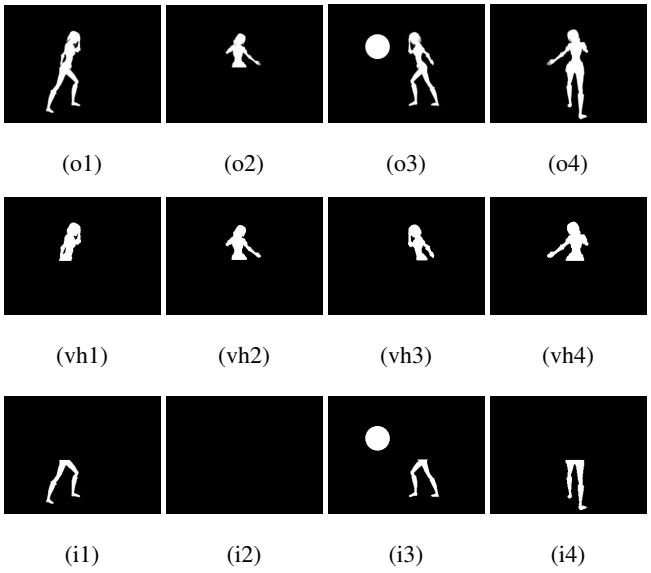

Fig. 1. The first row of images correspond to synthetic silhouettes where some errors have been intentionally introduced: In (o2), the bottom part of the silhouette has been deliberately removed and, in (o3), a false alarm has been incorporated. The second row of images shows the projection of the VH 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 IS are shown.

tic errors. In [7, 8] the epipolar tangency constraint (testing correspondences of the frontier points) is used as a necessary condition for shape consistency. However, the authors discard using the area of each silhouette that lies outside the visual hull for being slow and not suitable for pose estimation [7].

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 misclassification. Although our approach is voxel-based, we propose a general framework where any Projection Test can be used.

3. SHAPE FROM INCONSISTENT SILHOUETTE (*SFIS*)

Inconsistencies in the regions of the silhouettes can be detected by reconstructing the VH using SfS methods and projecting it back to examine how 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 Hull:IH*) which project to inconsistent regions in the silhouettes (*Inconsistent Silhouettes:IS*). Following, we formalize the concept of IH and IS and propose a procedure for estimating it.

3.1. Inconsistent Hull (*IH*)

The geometric concept of IH is introduced as the volume where there does not exist a shape which could possibly explain the observed silhouettes. Alternatively, the IH can be defined as the union of all the inconsistent cones, formed by the back-projection of the IS into the 3D scene. The IS are the resulting silhouettes after subtracting the original silhouettes with the projection of the visual hull (see Fig. 1 for an example using the Kung-Fu Girl dataset[1]). Thus, when the set of silhouettes is consistent then *all* the IS are empty, and the IH is also empty. However, when *a single* inconsistency appears in at least one silhouette then the IH will not be empty either.

From the two equivalent definitions above, it follows that the IH is disjoint of the VH ($VH \cap IH = \emptyset$). This can be observed in Figs. 2 and 3, where different situations have been depicted:

In Fig. 2(a), camera C misses the foreground object 1. The miss-detection entails an inconsistent set of silhouettes. However, the projections of object 2 are consistent, and therefore object 2 will be correctly reconstructed by standard SfS algorithms. Further inspection of the figure indicates that the IH in this case corresponds to the union of the visual cones $camA{\rightarrow}obj1$ and $camB{\rightarrow}obj1$. Moreover, the figure suggests that the more inconsistent cones intersect, the higher the chances of having missed an object in a certain camera. In Fig. 2(b), there have been 2 false region detections in camera B (α, β). The first one (α) does not intersect simultaneously with the rest of C-1 cones and therefore it does not produce any 3D errors. The other false alarm (β) leads to a false 3D object detection (marked in black) for intersecting simultaneously with C-1 visual cones. Furthermore, cone β is consistent, making the error undetectable with the consistency principle.

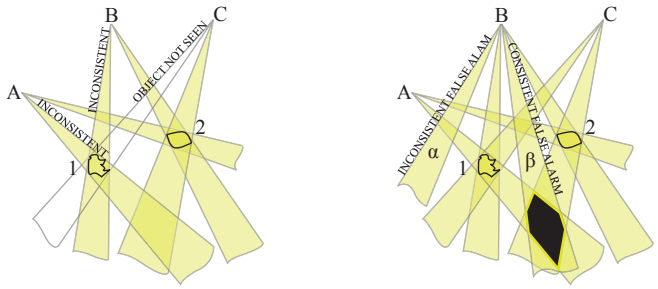

Fig. 2. In (a), object 2 is correctly detected in all the cameras, but object 1 is missed in camera C. The union of the cones $camB{\rightarrow}obj1$ and $camA{\rightarrow}obj1$ forms the IH. In (b) there have been two false region detections in camera B (α, β). The first one (α) forms an inconsistent cone, while the second one (β) forms a consistent one for intersecting with other C-1 visual cones.

Fig. 3 shows a slightly modified scenario. In this case, there is another object (object 3), which 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 indicates that the number of inconsistent cone intersections is not a sufficient condition for deciding whether there have been misses in some silhouettes or not. Furthermore, the figure also suggests that the number of occlusions (consistent *-not due to false alarms-* foreground projections) in the IH should be considered in any IH classification scheme.

In order to estimate the IH, we need to determine the unions of the inconsistent cones, similarly as SfS methods determine the intersections of the visual cones. We develop the concept of Shape from Inconsistent Silhouette using a voxel-based approach. The detailed process for the IH voxelization is shown in Algorithm 1.

In the voxel-based approach, the role of the inconsistent silhouettes (difference between *silhouettes* and *VH projection*) is re-

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

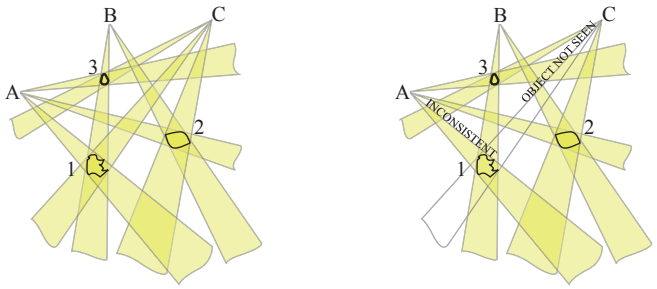

Fig. 3. In (a), objects 1, 2 and 3 are correctly detected in cameras A, B and C. In (b), object 1 is not reconstructed. The IH is smaller that its counterpart in Fig. 2 due to the occlusion of object 1.

placed by the difference between the Projection Test of the *silhouettes*: $PT_c(voxel, S(c))$ and the Projection Test of the *VH projection*: $PT_c(voxel, VH_{proj}(c))$. Note that the Projection Test can be any function designed to determine whether a voxel has been detected in a silhouette or not. For instance, a Projection Test may require all pixels where the voxel projects to be foreground, or just some of them[4]. In addition, any Projection Test will have an associated probability of miss $P(Miss_{2D})$ and false alarm $P(FA_{2D})$.

Algorithm 1 Voxelization of the IH

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**
  $IH(voxel) \leftarrow$ false
  for all c so that $PT_c(voxel, S(c))$ is true **do**
    if $PT_c(voxel, S(c)) \neq PT_c(voxel, VH_{proj}(c))$ **then**
      $IH(voxel) \leftarrow$ true

3.2. Unbiased Hull (UH)

The IH contains all the volumetric points which cannot explain 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 IH which is better explained as Shape for minimizing the probability of voxel miss-classification.

The classification of the voxels in the IH has to be optimal based on all the characteristics we can gather from them: First off, each voxel in the IH has an associated number of inconsistencies ($\mathcal{I}$), which corresponds to the number of inconsistent cone intersections in the voxel. A voxel 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. Finally, it can also be associated with the number of views where it projects to background ($\mathcal{B}$). For instance, voxels corresponding to object 1 in Fig. 3(b), have $\mathcal{I} = 1$, for being in the inconsistent cone $camA \rightarrow obj1$; $\mathcal{O} = 1$, for intersecting with the consistent occluding cone: $camB \rightarrow obj3$; and $\mathcal{B} = 1$, for being in the inexistent cone $camC \rightarrow obj1$.

From a practical point of view, $\mathcal{I}$ corresponds to the number of views where $PT_c(voxel, S(c)) = true \neq PT_c(voxel, VH_{proj}(c))$, $\mathcal{O}$ corresponds to the number of views where $PT_c(voxel, S(c)) = true = PT_c(voxel, VH_{proj}(c))$, and $\mathcal{B}$ to the number of views where $PT_c(voxel, S(c)) = false$, being $voxel \in IH$. Note that $\mathcal{I}$, $\mathcal{O}$ and $\mathcal{B}$ are such that $C = \mathcal{I} + \mathcal{O} + \mathcal{B}$, with C cameras.

Some further considerations regarding $\mathcal{I}$, $\mathcal{O}$ 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, the probability of having $C - \mathcal{I} - \mathcal{O}$ simultaneous misses increases. Contrarily, as $\mathcal{I}$ rises, the probability of having $\mathcal{I}$ simultaneous false alarms decreases. Based on this reasoning, there must exist an optimal threshold $\widehat{T}$ such that if $\mathcal{I} \geq \widehat{T}$, then the voxel is better explained as Shape (with $C - \mathcal{I} - \mathcal{O}$ misses) than Background (with $\mathcal{I}$ false alarms):

$$\begin{array}{lcl} \mathcal{I} \geq \widehat{T} & \Rightarrow & \text{decide } \textit{Shape} \\ \mathcal{I} < \widehat{T} & \Rightarrow & \text{decide } \textit{Background} \end{array} \quad (1)$$

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

$$\widehat{T} = \underset{T}{\operatorname{argmin}}\, P(Err_{3D}) \quad (2)$$

A voxel may be miss-classified if it is wrongly classified as Shape (false alarm) or if it is wrongly classified as Background (miss). The total classification error is then:

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

where P_B and P_S are the priors that a voxel forms part of the Background or Shape, respectively[2], and $P(FA_{3D})$ and $P(M_{3D})$ correspond to the probabilities of false alarm and miss in a voxel.

Let's first examine the probability of false alarm $P(FA_{3D})$:

$$P(FA_{3D}) = \sum_{i=max(T,1)}^{C-\mathcal{O}-1} \binom{C}{i} P(FA_{2D})^i (1 - P(FA_{2D}))^{C-i}, \quad (4)$$

corresponding to the summation of all the possible cases of false alarm in a voxel, between the limits explained next.

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. As the voxel forms part of the Background, then all the inconsistencies in the voxel ($\mathcal{I}$) correspond to false alarms of the Projection Test. Therefore, false alarms in the voxels are produced when the number of inconsistencies equals or surpasses the decision threshold T.

It has to be noted that the absolute minimum number of 2D false alarms which is possible in a Background voxel of the IH is 1, as 0 false alarms would correspond to a consistent voxel in the IH, which is an impossible situation. Thence, the lower limit is: $max(T, 1)$.

Analogously, the number of inconsistencies can be up to $C - \mathcal{O} - 1$. In terms of *consistency*, even though occlusions ($\mathcal{O}$) correspond to foreground projections, they cannot be considered as possible false alarms for having been validated by consistent voxels of the VH. Also note that the maximum number of false alarms cannot be $C - \mathcal{O}$, as this would correspond to a consistent foreground voxel in the IH, which is another impossible situation.

Thus, the resulting equation is the binomial of $P(FA_{2D})$ going from $max(T, 1)$ to $C - \mathcal{O} - 1$, where $P(FA_{2D})$ is the probability that the Projection Test is wrongly passed in a certain silhouette.

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

The opposite miss-classification case 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. A voxel is wrongly classified as Background if $\mathcal{I} < T$, which is equivalent to say that $\mathcal{B} \geq C - \mathcal{O} - T + 1$. Then, the probability of miss $P(M_{3D})$ in the IH, can be expressed in a similar manner as with false alarms:

$$P(M_{3D}) = \sum_{i=max(C-\mathcal{O}-T+1,1)}^{C-\mathcal{O}-1} \binom{C}{i} P(M_{2D})^i (1 - P(M_{2D}))^{C-i}, \quad (5)$$

where $P(M_{2D})$ corresponds to the probability that the Projection Test has not been passed by error.

Once that the probability of voxel miss-classification has been expressed, $\widehat{T}$ can be easily obtained by doing an exhaustive search of the minimum $P(Err_{3D})$ over all possible $T \in [0, C]$. Let us finally mention that $\widehat{T}$ can be obtained in $O(log(C))$ under certain convexity conditions. Details will be discussed in a future publication.

4. RESULTS

The theoretical benefits of *SfIS*, shown in Fig. 4, have been confirmed, using both synthetic and real data (collected in the smart room of the UPC). The system has been evaluated using 4 to 25 cameras, proving not to be sensitive with smooth variations of $P_{FA}(2D)$ and $P_M(2D)$. An aspect of interest of *SfIS* is that it behaves as traditional *SfS* when $P_{FA}(2D)$ is high or $P_M(2D)$ is low. In these cases, $\widehat{T} = C$, forcing the system to always decide Background.

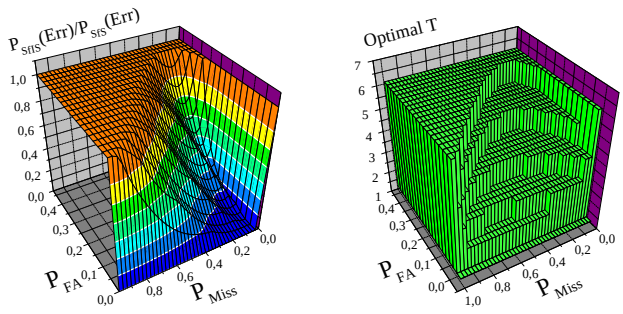

Fig. 4. The ratio $\frac{P_{SfIS}(Err_{3D})}{P_{SfS}(Err_{3D})}$, and $\widehat{T}$ for different values of $P_{FA}(2D)$ and $P_M(2D)$, with 6 cameras, $P_B = 0.9$ and $\mathcal{O} = 0$. $P_{SfS}(Err_{3D})$ is equivalent to $P_{SfIS}(Err_{3D})$, with $T = C$.

In Fig. 5, a real world scenario (with some additional false alarms in (o1) and (o4)) is shown. In this case the foreground segmentation has been done using [3]. In (o2), 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. Note that the miss-detection in (o2) has been propagated to the rest of silhouettes. The bottom row shows the projection of the $VH \cup UH$ in white and gray, respectively. The projection of the arm is recovered, even in (p2), while remaining unaffected to the artificial false alarms.

5. CONCLUSION

In this paper we have presented a novel scheme for effective Shape from Silhouette using sets of inconsistent silhouettes. The scheme exploits the consistency principle, and performs an error detection and correction procedure to recover the most probable consistent silhouettes. Experiments have demonstrated favorable results on various synthetic and real-world scenarios. Some of the future works include giving feedback to the background learning techniques to make more trustworthy background models.

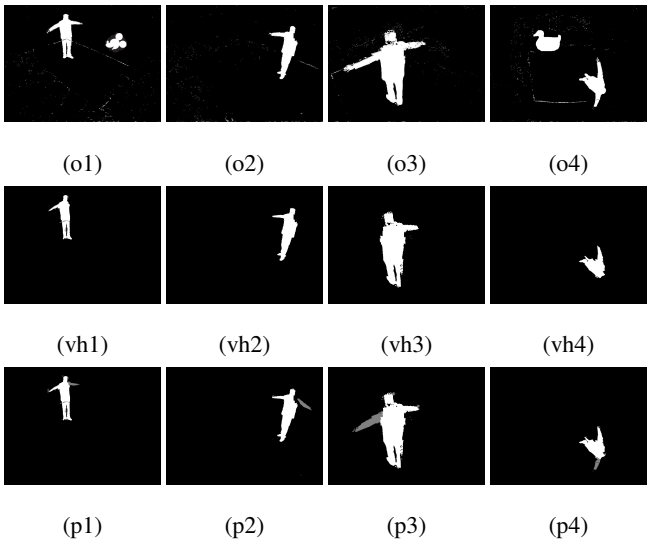

Fig. 5. Silhouettes, projection of the VH and projection of the $VH \cup UH$, in first, second and third row, respectively. The voxels' size has been chosen so that the projection of any voxel in the Shape is comprised within a pixel in all the silhouettes. Therefore, $P_{FA}(2D)$ and $P_M(2D)$ concur with the probabilities of FA and Miss of the background learning technique. In this case we have used $P_{FA}(2D) = P_M(2D) = 0.1$, and P_B has been selected based on the percentage of voxel occupancy in the VH.

6. REFERENCES

[1] Bruce Guenther Baumgart, *Geometric modeling for computer vision.*, Ph.D. thesis, 1974. 1

[2] A. Laurentini, "The visual hull: A new tool for contour-based image understanding," *Proc. Seventh Scandinavian Comperence on Image Processing*, pp. 993–1002, 1991. 1

[3] Chris Stauffer and W. Eric L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. on Pattern Anal. and Machine Intel.*, vol. 22, no. 8, pp. 747–757, 2000. 1, 4

[4] 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 the 2000 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '00)*, June 2000, vol. 2, pp. 714 – 720. 1, 3

[5] D. Snow, P. Viola, and R. Zabih, "Exact voxel occupancy with graph cuts," in *Proc. of the 2000 IEEE Conf. on Computer Vision and Pattern Recog. (CVPR '00)*, 2000, pp. 345–353. 1

[6] J.L. Landabaso and M. Pardas, "Foreground regions extraction and characterization towards real-time object tracking," in *Proceedings of Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI '05)*, 2005. 1

[7] K. Forbes, A. Voigt, and N. Bodika, "Using silhouette consistency constraints to build 3d models," PRASA, 2003. 2

[8] K.-Y. K. Wong, *Structure and Motion from Silhouettes*, Ph.D. thesis, Dept. of Engineering, Univ. of Cambridge, 2001. 2