Using Silhouette Consistency Constraints to Build 3D Models

Keith Forbes[1] Anthon Voigt[2] Ndimi Bodika[2]

[1]Digital Image Processing Group
Department of Electrical Engineering
University of Cape Town
Private Bag, Rondebosch, 7701
kforbes@dip.ee.uct.ac.za

[2]Automation and Informatics Group
TSS Technology
De Beers Consolidated Mines
P. O. Box 82851, Southdale, 2135
{Anthon.Voigt, Ndimi.Bodika}@debeersgroup.com

Abstract

The visual hull is the largest object that is consistent with a set of silhouette views of an actual object. The visual hull can be built from a set of silhouettes in which the viewpoint corresponding to each silhouette is known. We show how several sets corresponding to the same rigid object, each containing a small number of silhouettes, can be merged into a single large set. In order to do this, the relative pose between the sets must be computed so that each viewpoint can be specified in a common reference frame. We show how the poses can be computed by enforcing silhouette consistency constraints between the sets. The single merged silhouette set can then be used to build a visual hull model that is a closer approximation to the actual object than visual hulls built from any of the original sets.

1 Introduction

Shape-from-silhouette techniques are often used as a relatively simple means for forming approximate 3D models of an object. The visual hull is the largest object that is consistent with a given set of silhouettes and associated known viewpoints; it is often used as an approximate 3D model.

1.1 The Visual Hull Concept

The term visual hull was coined by Laurentini [7] in the 1990s, but the use of the largest silhouette-consistent object as a means for 3D modelling dates back to the work of Baumgart in the 1970s [2]. Initially, the term visual hull was used to describe the largest object consistent with all possible silhouettes, but the term is usually used to refer to the largest object that is consistent with a finite set of available silhouettes.

The visual hull concept is illustrated in Figure 1. Figure 1(a) shows two silhouette views of a duck (the actual object that is being modelled). Camera centres are represented by small spheres. For convenience, the image planes are placed *in front* of the camera centres, and the projected silhouette views are shown non-inverted; for the purposes of this work, this setup is geometrically equivalent to placing the image planes behind the camera centres. *Visual cones* corresponding to each silhouette are shown in Figure 1(b). A visual cone is the volume of space that the actual object cannot lie outside of, given the observed silhouette. The intersection of the visual cones is the visual hull (shown in Figure 1(c)). The visual hull cannot be smaller than the actual object. With two silhouettes, the visual hull is often a poor approximation to the actual object. However, if further silhouette views are added, more information about which volumes of space are empty is added, and the visual hull becomes a better approximation to the actual object.

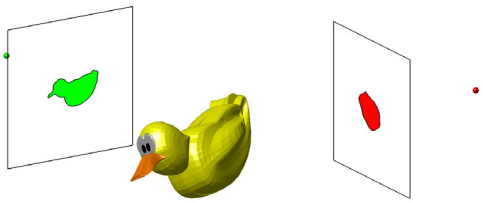
In order to determine the visual hull corresponding to a set of silhouettes, the cameras that produced the images must be calibrated. This means that the internal camera parameters (such as focal length, principal point and lens distortion parameters) and the pose or external camera parameters (the position and orientation of the cameras in a common reference frame) must be known.

1.2 Related Work

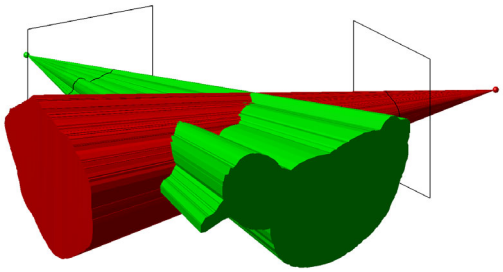
Various means have been used to determine the pose corresponding to each silhouette view. For instance, Shakhnarovich *et al.* [11] describe a system that uses four fixed cameras. The pose of each camera is determined once-off (in a common reference frame) using a calibration object. Okatani and Deguchi [10] use a gyro sensor to determine the orientation of a single camera that is moved to different viewpoints. The positional component of the pose is then determined from the silhouette images themselves. Niem and Buschmann [9] use a single camera and a turntable to obtain multiple silhouette images of an object at different viewpoints. The relative poses of the turntable platform with respect to the camera are determined using a calibration object. Wong [12] also uses turntable sequences, but is able to determine the relative poses of the silhouettes without the use of a calibration object; the silhouettes of the object being modelled, along with the assumption of circular motion, are used to determine the relative poses. The solution is subsequently refined by removing the assumption of perfect circular motion and adjusting the pose estimates to minimise a cost function based on the epipolar tangency constraint (explained in Section 2).

1.3 Overview

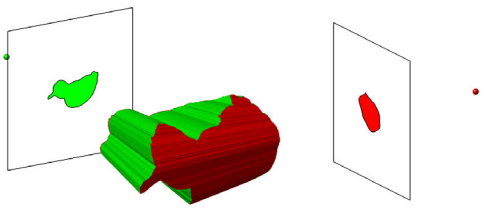
In this work, a method for obtaining a large set of silhouettes using a system consisting of a small number of fixed cameras is presented. The relative poses of the cameras are known in a common reference frame (a once-off calibration process is used). Our sys-



(a)



(b)



(c)

Figure 1: Two silhouette views of a duck showing (a) the cameras, each represented by a camera centre and image plane, (b) the visual cones corresponding to each of the two silhouettes, and (c) the visual hull corresponding to the two silhouettes.

tem captures sets of silhouettes* of an object. Each set of silhouettes is captured with the object in a different pose, so that each set consists of silhouettes from different viewpoints. We show that for a rigid object, the relative poses of the object can be determined by minimising a cost function based on a measure of inconsistency between the different sets of silhouettes. (A set of silhouettes is consistent if a 3D object exists that could have produced the set.) The cost for a candidate relative pose is a measure of the inconsistency that would occur across the silhouettes of two sets, if the candidate pose were used to merge the sets by specifying all viewpoints in a common reference frame. Once the relative poses are computed, the sets of silhouettes can all be merged into a single large set of silhouettes. A visual hull model can then be built from the large set of silhouettes. This visual hull model is a closer approximation to the actual object than a visual hull model built from any of the original sets. This is because a larger number of silhouettes with known viewpoints allows a greater portion of space known not to belong to the actual object, to be removed. The method allows a greater range of viewpoints to be used than can be obtained from a single camera and turntable setup. For instance, a turntable setup would not be able to model the hole of a doughnut, if the doughnut were placed flat on the turntable. This is because the viewpoints corresponding to a turntable sequence of silhouettes lie on a circle in the viewing sphere, whereas with our method, the viewpoints are well-distributed over the entire viewing sphere.

Section 2 describes the conditions that a silhouette set must fulfil in order to be consistent. Because of noise associated with the calibration parameters and segmentation, silhouettes will not, in practice, be perfectly consistent. Methods for computing a *degree* of inconsistency are therefore presented. Section 3 describes how the relative pose between sets of silhouettes can be determined by minimising the degree of inconsistency between them. Some results obtained using our five camera system are shown in Section 4. Section 5 summarises the paper.

2 Silhouette Consistency

There are two constraints which limit the possible shape of a silhouette at a given viewpoint. The possible shape is limited by the information contained in the remaining silhouettes of the set. First, we describe the *epipolar tangency constraint*, which is used in this work. Our formulation of the cost function is very similar to that of Wong [12]. Next, for completeness, we briefly describe another constraint that we call the *visual hull projection constraint*. We do not currently make use of this constraint for pose estimation.

2.1 The Epipolar Tangency Constraint

This section introduces several geometric concepts such as epipoles, frontier points, epipolar tangencies and the essential matrix. The epipolar tangency constraint, which is based on these concepts, is then explained. It is shown how a measure of inconsistency, based on the epipolar tangency constraint, can be computed. If a candidate pose is used to merge two sets of silhouettes, then the measure of inconsistency can be used as a cost associated with the candidate pose. By adjusting the pose parameters to minimise the cost, a good estimate of the true pose can be computed.

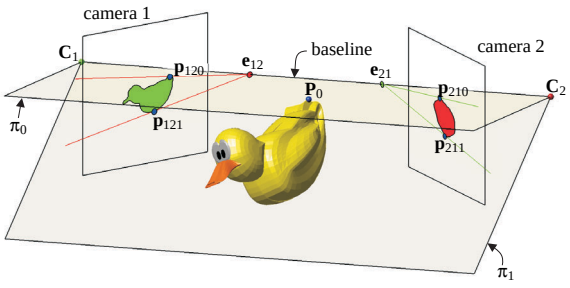
Figure 2 shows the same scene as shown in Figure 1, along with some additional points and planes. The line joining the two camera centres $\mathbf{C}_1$ and $\mathbf{C}_2$ is called the baseline. It pierces the image plane of camera 1 at $\mathbf{e}_{12}$ and the image plane of camera 2 at $\mathbf{e}_{21}$. The projection of a camera centre onto the image plane of another camera is termed an *epipole*. The points $\mathbf{e}_{12}$ and $\mathbf{e}_{21}$

*In this paper, a set of silhouettes refers to silhouettes whose viewpoints are known in a common reference frame.

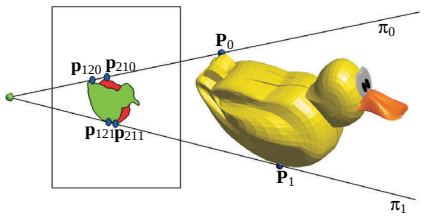
are epipoles. In the figure, the epipoles are represented as small circles (projections of spheres) on the image planes. Since we are not looking directly onto the image planes, the circles appear as ellipses. Note that the epipoles do not necessarily lie in the visible region of the image plane (corresponding to the finite extent of the camera's sensor).

The two planes π_0 and π_1 that pass through the baseline and are tangent to the duck are shown. As long as the baseline does not pass through the object, there will be two such planes for any object. The points $\mathbf{P}_0$ and $\mathbf{P}_1$, where the planes touch the object's surface, are termed *frontier points* [5]. Since the planes pass through both camera centres and graze the surface of the object, the frontier points project onto the silhouette boundary in both views. The projection of a frontier point is the tangency point of a silhouette tangent line that passes through the epipole. A projection of a frontier point is therefore termed an *epipolar tangency*. The epipolar tangencies $\mathbf{p}_{120}$ and $\mathbf{p}_{210}$ are projections of $\mathbf{P}_0$, and $\mathbf{p}_{121}$ and $\mathbf{p}_{211}$ are projections of $\mathbf{P}_1$. (The notation $\mathbf{p}_{ijk}$ is used so that i indicates the number of the camera whose image plane the point lies on, j indicates the number of the opposite camera of the silhouette pair, and k indicates to which of the two frontier points $\mathbf{p}_{ijk}$ corresponds.)

For the purpose of computation, silhouettes are represented as polygons in normalised image coordinates. The normalised coordinates for an image point are the x- and y-coordinates of the position where the corresponding 3D ray pierces the $z = 1$ plane in the camera's reference frame. Since the internal camera parameters are known (calibrated cameras are used), pixel coordinates can be converted to normalised image coordinates.

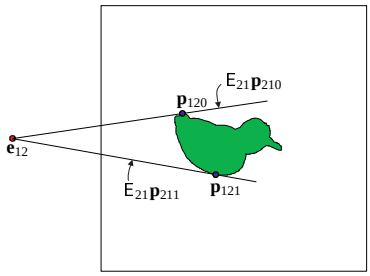


(a)

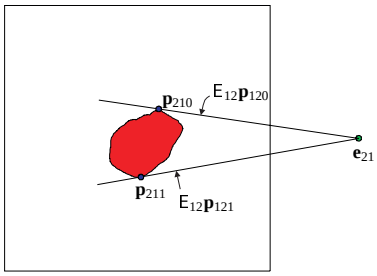


(b)

Figure 2: Two views of the epipolar geometry of a scene: (a) shows a front view, and (b) shows a side view looking onto the scene in a direction parallel to the baseline.



(a)



(b)

Figure 3: The epipolar tangency constraint: the epipolar tangent line touches the silhouette at the projection of the frontier point, as shown in (a) and (b); the projection of this line onto the image plane of the opposite camera is constrained to coincide with the opposite epipolar tangency line.

The intrinsic geometry between the views i and j can be encapsulated by the 3×3 *essential matrix* E_{ji} [6]. If $\mathbf{x}_i$ represents the homogeneous coordinates of an image point in normalised image coordinates from view i, and $\mathbf{x}_j$ represents the corresponding point in view j, then $\mathbf{x}_i$ is constrained to lie on the line $\mathsf{E}_{ji}\mathbf{x}_j$ in view i so that

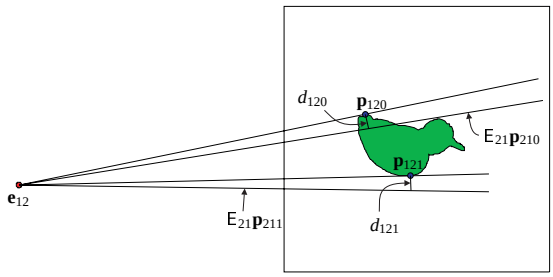
$$\mathbf{x}_i^{\mathsf{T}} \mathsf{E}_{ji} \mathbf{x}_j = 0. \tag{1}$$

If the relative pose between view i and view j is described by a rotation represented by the matrix R followed by a translation represented by the vector $\mathbf{t}$ that transform points from the reference frame of camera j to the reference frame of camera i, then the essential matrix can be computed using

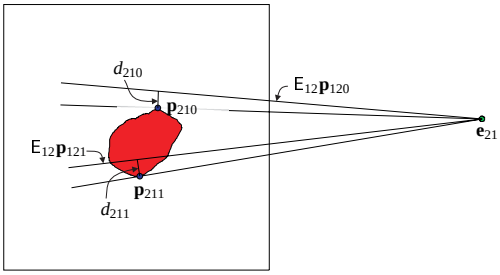
$$\mathsf{E}_{ji} = [\mathbf{t}]_\times \mathsf{R}. \tag{2}$$

The antisymmetric matrix $[\mathbf{t}]_\times$ is computed from the translation vector $\mathbf{t} = [t_x, t_y, t_z]^{\mathsf{T}}$ using

$$[\mathbf{t}]_\times = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix}. \tag{3}$$



(a)



(b)

Figure 4: Epipolar tangent lines with the projection of the epipolar tangent lines of the opposite view and incorrect pose information: since the pose information is incorrect, the epipolar tangent lines do not project onto one another. The silhouettes are inconsistent with one another for the given viewpoints. The reprojection error is a measure of the degree of inconsistency.

The essential matrix can therefore easily be computed for a given pose.

The essential matrix can be used to compute the normalised image coordinates of the epipoles. The epipoles $\mathbf{e}_{ij}$ and $\mathbf{e}_{ji}$ are the right and left nullspaces of E_{ji} respectively. If $\mathsf{U}\Sigma\mathsf{V}^\mathsf{T}$ is the singular value decomposition of E_{ji}, then the last column of V is $\mathbf{e}_{ij}$ (in homogeneous coordinates). Since $\mathsf{E}_{ij} = \mathsf{E}_{ji}^\mathsf{T}$, the epipole $\mathbf{e}_{ji}$ can be computed in a similar manner.

Figures 3(a) and (b) show the epipolar tangency lines for each silhouette image of the duck example. Each line lies in a tangent plane containing a frontier point, and therefore must project onto the corresponding line in the opposite image: this is the epipolar tangency constraint. In other words, in the noiseless case, the line passing through $\mathbf{e}_{ij}$ and $\mathbf{p}_{ijk}$ is the same line as $\mathsf{E}_{ji}\mathbf{p}_{jik}$.

If there are inaccuracies in the silhouettes or the pose, then the line passing through $\mathbf{e}_{ij}$ and $\mathbf{p}_{ijk}$ will not, in general, be the same line as $\mathsf{E}_{ji}\mathbf{p}_{jik}$. Figures 4(a) and (b) show the noisy case in which there are inaccuracies in the assumed relative pose between the cameras. Note that the epipoles are positioned differently to Figure 3, since the pose is incorrect. The projection of the opposite camera's epipolar tangency line is not exactly coincident with the epipolar tangency line on the image plane. Reprojection errors can be computed as a measure of the inconsistency between a pair of silhouettes with an associated pose value. The reprojection error is the shortest distance from an epipolar tangency to the epipolar line of the corresponding point in the opposite image. The figure shows the reprojection errors d_{120}, d_{121}, d_{210} and d_{211}.

The distance d_{ijk} between an epipolar tangency $\mathbf{p}_{ijk}$ and the projection of the epipolar line from the opposite camera that passes through the tangency point $\mathbf{p}_{jik}$ can be computed using the essential matrix.

$$d_{ijk} = \frac{\mathbf{p}_{ijk}^\mathsf{T}\mathsf{E}_{ij}\mathbf{p}_{jik}}{\sqrt{(\mathsf{E}_{ij}\mathbf{p}_{jik})_1^2 + (\mathsf{E}_{ij}\mathbf{p}_{jik})_2^2}} \tag{4}$$

The expressions $(\mathsf{E}_{ij}\mathbf{p}_{jik})_1^2$ and $(\mathsf{E}_{ij}\mathbf{p}_{jik})_2^2$ denote the first and second elements of the vector $(\mathsf{E}_{ij}\mathbf{p}_{jik})^2$. Note that $\mathbf{p}_{ij0}$ and $\mathbf{p}_{ij1}$ are vertices of the polygon representing the silhouette. To determine which two polygon vertices are the epipolar tangencies, the slope of the line from each polygon vertex to the epipole must be examined.

For two sets of silhouettes and an associated pose, a measure of inconsistency across all silhouettes can be computed by considering all possible pairings of silhouettes in the first set with silhouettes in the second set. This measure of inconsistency is treated as the cost function associated with the pose. For a first set consisting of m silhouettes, and a second set consisting of n silhouettes, it is computed by summing the squared reprojection errors across all cases as follows:

$$\text{cost} = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=0}^{1} d_{ijk}^2. \tag{5}$$

2.2 The Visual Hull Projection Constraint

We note that the fulfilment of the condition imposed by the epipolar tangency constraint is a necessary, but not a sufficient condition for silhouette consistency.

A silhouette cannot lie outside the projection of the visual hull of the remaining silhouettes in the set. If it were to do so, it would be providing contradictory information, since the remaining silhouettes indicate that the volume of space corresponding to the area outside the visual hull projection is empty. A measure of this form of inconsistency could be made by computing the area of each silhouette that lies outside the visual hull projection formed from the remaining silhouettes. This measure would be slow to compute and is thus not suited for pose optimisation, since the measure would have to be computed for each set of pose parameters evaluated by the optimisation routine.

3 Pose Estimation

In order to merge two silhouette sets into a single set, the relative pose between the two sets must be computed. This pose corresponds to the difference in pose between the two poses of the actual object on the two occasions that multiple silhouettes of the object were captured. The pose is determined by adjusting the pose parameters to minimise the cost function of Equation (5).

A quaternion and a translation vector are used to parameterise the pose between two sets. The Levenberg-Marquardt method [6] is used to adjust the seven pose parameters to minimise the cost function. It is important to provide the Levenberg-Marquardt routine with a good initial estimate of the pose parameters, otherwise

the routine may converge to a local minimum that is far from the global minimum.

Wong [12] minimises essentially the same cost function, but assumes circular motion between individual silhouettes in order to to form an initial estimate. Additional silhouettes from arbitrary viewpoints can later be incorporated, but the user must manually determine the initial pose estimate for each additional view.

Unlike Wong, we are dealing with sets of silhouettes, rather than individual silhouettes; in our case the relative poses between silhouettes in the set is known in advance. This means that the visual hulls corresponding to the sets can be used to form initial pose estimates.

The translational component of the pose is estimated by using the difference between the centroids of the visual hull models corresponding to each of the two sets. The rotational component can be estimated by determining the transformation that will align the principal axes of the two visual hull models. Since there are four ways in which the principal axes can be aligned, we use all four as initial estimates for four separate optimisations. If none of these four starting points yield a pose with sufficiently low associated cost, then random rotations drawn from a uniform distribution [1] are computed and used as initial estimates until a sufficiently low associated cost is obtained. This approach is possible since the cost function is fast to evaluate.

Once two silhouette sets have been merged into one, additional sets can be merged with the merged set in the same manner.

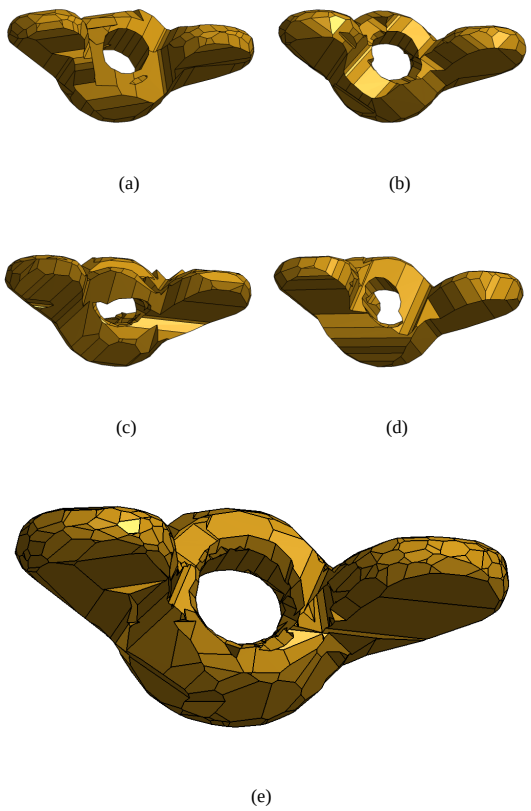


Figure 5: Visual hull models of a wing nut: (a)–(d) show four models each built from five silhouettes, (e) shows the model built from the 20 silhouettes used in (a)–(d) after the poses of all silhouettes have been determined in a common reference frame.

4 Experimental Results

Our system consists of five CCD cameras that are mounted so that they are well-spaced about the viewing hemisphere. The five image set is taken of the object while it is in free flight. To obtain multiple sets of five images, we simply drop the object into the common field of view several times. Typically, the orientation of the object varies substantially each time it is dropped past the cameras.

We calibrate our cameras using an icosahedron calibration object with coded target patterns on its faces [4]. By locating the positions of coded target centres across multiple images, it is possible to infer the camera parameters.

A fast, simple threshold-based segmentation algorithm is used to determine the polygon that separates the foreground from the background. The number of vertices in the polygons is then reduced using the Douglas-Peucker method [3]. We have coded a C^{++} implementation of an efficient polyhedral visual hull algorithm [8] to build the polyhedral models. The polyhedral visual hull models can be viewed as VRML (Virtual Reality Modelling Language) models.

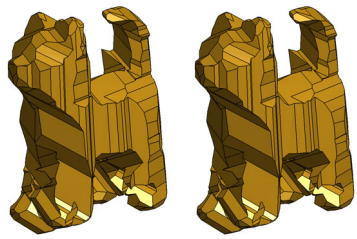
Figure 5 shows an example of visual hull models of a wing nut that was dropped through the system four times. Notice how each of the five-view models is quite coarse, with a lot of extra volume around the hole region, yet in the twenty-view case, the combined information can be used to build a visual hull model that looks relatively accurate.

Figure 6 shows an example of a toy cat. Again, notice how the twenty-view model looks more accurate than any of the visual hull models built from the five silhouettes in the original sets. Figure 7 shows the twenty visual cones that correspond to the twenty silhouettes and that were used to build the model shown in Figure 6. Notice how the cones are well-distributed about the viewing sphere.

The entire process of determining the relative poses (three poses must be determined to merge four silhouette sets) and computing the twenty-view visual hull model takes approximately 3 seconds on a 2.4 GHz Pentium 4 machine for the examples shown. Each of the coarse five-view visual hull models is built from the silhouette polygons in approximately 70 ms.

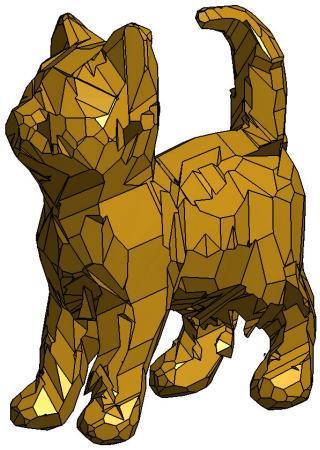
5 Summary

We have shown that sets of silhouettes of a rigid object can be merged into a single silhouette set by determining the relative pose between sets. The relative pose is determined by minimising a cost function based on the epipolar tangency constraint. The minimisation routine requires a good initial pose estimate in order to be likely to converge to the correct set of pose parameters. Wong [12] showed that a good initial estimate could be made for the pose of individual silhouettes, if it could be assumed that there is a circular motion between successive views. In this paper, we have shown that a good initial pose estimate can also be obtained if sets of



(a) (b)

(c) (d)



(e)

Figure 6: Visual hull models of a toy cat: (a)–(d) show four models each built from five silhouettes, (e) shows the model built from the 20 silhouettes used in (a)–(d) after the poses of all silhouettes have been determined in a common reference frame.

small numbers of silhouettes are used. The relative viewpoint of each silhouette in a set is known in advance. An initial estimate of the translation component of the pose is computed as the difference between the centroids of the two visual hull models corresponding to the two sets that are to be merged. An initial estimate of the rotational component of the pose can either be computed by determining the alignment between the principal axes of the visual hulls, or by testing successive random rotations. The approach of

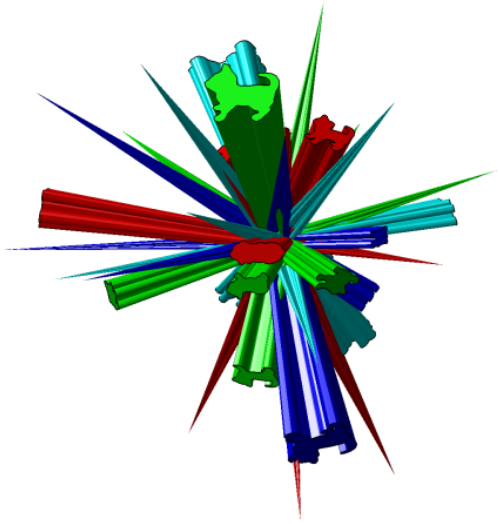


Figure 7: The twenty visual cones of the cat

using random rotations is feasible because the cost function is fast to compute for a given pose parameter set, and because the number of degrees of freedom for a rotation is relatively small. Our results show that the visual quality of 3D models built from merged silhouette sets is better than that of the models built from the original silhouette sets.

References

[1] James Arvo. Fast random rotation matrices. In David Kirk, editor, *Graphics Gems III*, pages 117–120. Academic Press, 1992.

[2] Bruce G. Baumgart. *Geometric Modeling for Computer Vision*. PhD thesis, Stanford University, 1974.

[3] D. H. Douglas and T. K. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Canadian Cartographer*, 10(2):112–122, 1973.

[4] Keith Forbes, Anthon Voigt, and Ndimi Bodika. An inexpensive, automatic and accurate camera calibration method. In *Proceedings of the Thirteenth Annual South African Workshop on Pattern Recognition*. PRASA, 2002.

[5] Peter J. Giblin and Richard S. Weiss. Epipolar curves on surfaces. In *1994 ARPA Image Understanding Workshop*, volume II, November 1994.

[6] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[7] Aldo Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):150–162, February 1994.

[8] Wojciech Matusik, Chris Buehler, and Leonard McMillan. Polyhedral visual hulls for real-time rendering. In *Proceedings of Twelfth Eurographics Workshop on Rendering*, pages 115–125, June 2001.

[9] W. Niem and R. Buschmann. Automatic modelling of 3D natural objects from multiple views. In *European Workshop on Combined Real and Synthetic Image Processing for Broadcast and Video Production*, Hamburg, Germany, November 1994.

[10] Takayuki Okatani and Koichiro Deguchi. Recovering camera motion from image sequence based on registration of silhouette cones: Shape from silhouette using a mobile camera with a gyro sensor. In *Proceedings of the 6th IAPR Workshop on Machine Vision Applications MVA2000*, pages 451–454, 2000.

[11] G. Shakhnarovich, L. Lee, and T. Darrell. Integrated face and gait recognition from multiple views. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001.

[12] K.-Y. K. Wong. *Structure and Motion from Silhouettes*. PhD thesis, University of Cambridge, October 2001.