Sensing and Describing 3-D Structure

Peter K. Allen

Department of Computer Science
Columbia University
New York, NY 10027

Abstract

Discovering the three dimensional structure of an object is important for a variety of robot tasks. Single sensor systems such as machine vision systems cannot reliably compute three dimensional structure in unconstrained environments. Active, exploratory tactile sensing can be used to complement passive stereo vision data to derive robust surface and feature descriptions of objects. The control for tactile sensing is provided by the vision system which provides regions of interest that the tactile system can explore. The descriptions of surfaces and features are accurate and can be used in a later matching phase against a model data base of objects to identify the object and its position and orientation in space.

1. INTRODUCTION

Robotic systems are being built to perform complex tasks such as object recognition, grasping, manipulation and assembly. A common thread in all these tasks is the ability to understand the three dimensional geometric and topological structure of the objects involved. The first step in discovering the underlying three dimensional structure of an object through sensing is to compute depth and orientation at each point on the object, or what has been termed by Marr [10] as a "2½ D" sketch. Currently, there are several sensing systems that can derive depth from a scene. Among these are laser range finders [17, 9, 1], photometric stereo [7] and binocular stereo [3]. Determining which sensor to use is chiefly determined by the task domain. Laser imaging is potentially hazardous and has difficulty with shiny metal reflective surfaces. At present, it is a more expensive depth sensing technology than the other methods mentioned above. Photometric stereo puts great demands on the illumination in the scene and on properly understanding the reflectance properties of the objects to be viewed. In choosing a system to sense depth in general, unconstrained environments, binocular stereo has the advantage of low cost and ability to perform over a wide range of illuminations and object domains. It is also a well understood and simple ranging method, which motivates its use in a generalized robotics environment where many different task and object domains may be in effect. However, used as a single robotics sensing system, stereo has clear deficiencies. If there is a lack of detail on the object, only sparse depth measurements are possible. If too much detail is present, the matching process between image events can easily become confused. Detail also causes a marked degradation in performance as the potential match space increases. To overcome the limitations of such a sensor system a useful approach is to use multiple sensors. Multiple sensors can be used in a complementary fashion to extract more information from an environment than a single sensor [16, 12]. Tactile sensing is a good choice for a complementary sensor to vision in a generalized robotics environment for a number of reasons. Touch is able to directly measure the properties of objects we desire: their position and surface orientation. It also can sense visually occluded areas of a scene, making it more useful than the other ranging devices mentioned. Lastly, it is a low cost sensor that is required for tasks such as grasping and manipulation, making it a necessary part of a robotic system.

A system using passive stereo vision and active exploratory tactile sensing to recognize common kitchen items such as mugs, bowls, pitchers and plates has been built and is described in detail in [2]. A key component of this system are modules that create surface and feature descriptions of objects. This paper describes the procedures used to generate the surface and feature descriptions and reports results achieved with the method for a number of real objects.

The experimental hardware is shown in figure 1. The objects to be recognized are rigidly placed on the worktable and imaged by a pair of CCD cameras. The tactile sensor is mounted on a 6 degree of freedom PUMA manipulator that receives feedback from the tactile sensor allowing it to move across the surfaces of objects reporting contacts. The sections below describe how the vision and touch are integrated into robust surface and feature descriptions.

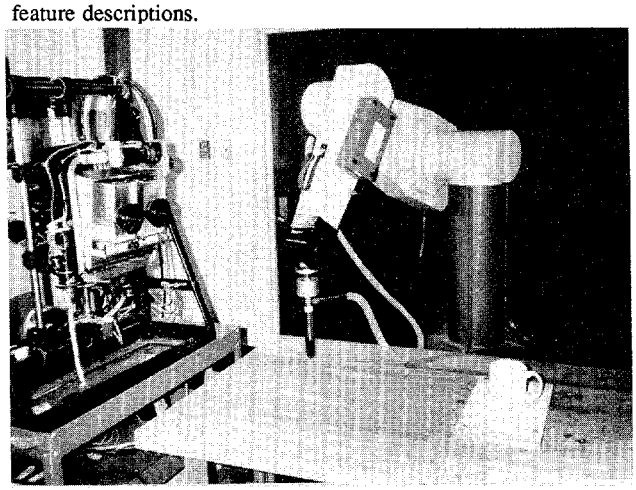


Figure 1. Experimental hardware.

2. VISION PROCESSING

Because we are using touch to supplement the vision data, we are not concerned with trying to extract as much information out of vision as possible. Rather than pushing vision processing to its limits with the resulting inevitable error, we choose to use only that part of vision that can be shown to be reliable and accurate. While this has the negative effect of producing sparse depth and surface information from vision, this is offset by the resulting high confidence and accuracy of the depth information produced from vision.

Depth is determined from stereo using a pair of CCD cameras which are aligned so that the epipolar lines lie on scan lines of the cameras, reducing the search space for matches. The Marr-Hildreth edge operator [11] is applied on both images and zero-crossings of the convolved images are found. The zero-crossings are then thresholded and closed contour regions are grown, yielding chains of connected contour pixels for each closed region. The stereo matching is based upon these closed contour chains. Pixels in each image are matched if they satisfy the following criteria:

1) The zero-crossings are on the same scan line.
2) The zero-crossing must belong to a closed contour of a grown region.
3) The zero-crossings must have an orientation at least 20° from horizontal.
4) The zero-crossings must have the same contrast sign.
5) The zero-crossings have the same orientation (within 30°).

If these criteria are met, then we have a set of candidate match pixels. These candidate matches are then further constrained by performing a correlation centered around each match, using a window size that is related to the size of the initial filter used in finding the zero-crossings in each image. By establishing high confidence levels (above 95%) for these correlations, only those matches that are robust will survive. The output of this matching is a sparse set of three dimensional points located on the boundary contours of closed regions in the image (figure 2). These regions represent smooth areas on the object since they have no zero-crossings in the interior. However, it can not be determined from vision alone if these regions are surfaces, cavities or holes. The tactile sensor can explore each region to determine its actual structure, guided by the sparse 3-D data computed at the regions boundaries.

3. TACTILE SENSING

Tactile sensing is a relatively new and underutilized sensing modality. Previous work in tactile sensing for recognition tasks has emphasised traditional pattern recognition paradigms on arrays of sensor data, similar to early machine vision work [6, 13, 14, 8]. Most sensing has been static in that the sensor is larger than the object and a single "touch" is used for recognition. Very little has been done on dynamic sensing and integrating multiple "touch frames" into a single view of an object.

Touch is different from vision in that is an active, exploratory sensing modality. Active touch sensing provides powerful shape information but it extracts its price for this information in demanding powerful control of the medium that makes it difficult to use. Blind groping on a surface with a tactile sensor is a poor and inefficient way of understanding three dimensional structure. Touch needs to be guided to be useful, and the vision data can provide this guidance to an active touch sensor.

The experimental tactile sensor used in this research was developed at L.A.A.S in Toulouse, France. It consists of a rigid plastic core covered with 133 conducting surfaces that is roughly the size and shape of a human index finger. The geometry of the sensor is an octagonal cylinder of length 228 mm. and radius 20 mm. On each of the eight sides of the cylinder there are 16 equally spaced conducting surfaces. The tip of the sensor contains one conducting surface, and there are four other sensors located on alternate tapered sides leading to the tip. The conducting surfaces are covered by a conductive elastomeric foam. The sensor is connected to a A/D converter that outputs the readings on all sensors in an eight bit gray value and the entire array of sensors may be read in a few milliseconds.

The organization of tactile sensing is on three distinct hardware and software levels. The low level is a series of programs that condition and sample the data coming from the sensor. The intermediate level consists of programs that move the robotic arm based upon feedback from the tactile sensor. The high level is used to provide information about the regions in space that are to be explored with the sensor. Algorithms exist to explore a region in space and determine if it is a surface, hole or cavity. Once a region is identified, it can be further explored by surface following algorithms that report contact points on surfaces and boundary contours of holes and cavities to a controlling host process. These contacts can be integrated with the 3-D contours from vision to build robust surface and feature descriptions.

4. EXPLORING REGIONS

The high level tactile processing will determine a region to explore by touch. Once a region is chosen to be explored, the intermediate level exploration program is invoked. This program will establish if the region discovered by the vision algorithms is a surface, hole or cavity. The program needs as input an approach vector towards the region. The orientation of the sensor is computed by calculating the least square plane P_{lsq} with unit normal $\mathbf{N}_{lsq}$ from the matched 3D stereo points that form the contour of the region. $\mathbf{N}_{lsq}$ then becomes the approach vector for the sensor. The arm control routines will orient the arm so that the tactile sensor's long axis is aligned with $\mathbf{N}_{lsq}$, pointing in the direction of the region's centroid as determined from the vision processing.

The arm is then moved along the sensor's long axis until contact with a surface or it moves beyond plane P_{lsq}, implying the presence of a hole or a cavity. If the sensor is able to travel its full length beyond P_{lsq} without contact, then a hole has been found. If it travels beyond a specified cavity threshold T_{cav} before contact, then it is a cavity. If the region is a surface, a surface patch description will be computed. If it is a hole or cavity, a boundary curve will be traced.

5. BUILDING SURFACE DESCRIPTIONS

At the lowest levels of recognition are the actual sensor primitives which can be pixels from vision or contact points from touch. This data is too lacking in structure to be useful by itself

and needs to be grouped into larger tokens. Surfaces are a natural choice for this next level of recognition because they are the components that vision sensors see and touch sensors feel. They have the added attributes of being more stable descriptions than points and are an effective means of compressing and abstracting point data. The particular form of bicubic surface patch that is being used in this research is a Coons'patch [4] which has been used extensively in computer graphics and computer aided design. The patches are constructive in that they are built up from known data and are interpolants of sets of three dimensional data defined on a rectangular parametric mesh. This gives them the advantage of axis independence, which is important in synthesizing these patches from sensory data. Being interpolating patches, they are able to be built from sparse data. The most important property possessed by these patches is their ability to form composite surfaces with C^2 (curvature continuous) continuity. The object domain (mugs, bowls, pitchers, plates) contains many curved surfaces which are difficult or impossible to represent using polygonal networks or quadric surfaces. A bicubic patch is the lowest order patch that can contain twisted space curves on its boundaries.

These patches define composite surfaces that can be made up of one or many curvature continuous patches. Level 0 surfaces are surfaces comprised of a single surface patch. They are defined on 2 x 2 rectangular knot set (figure 3). The information needed to build a level 0 surface is the 4 knot points, the tangents in each of the parametric directions and the twist vectors at these knots. There are two considerations in choosing the knot points. The points should be chosen at points of high curvature on the boundary curve and the knots need to be spaced uniformly in each of the parametric directions. The algorithm for choosing points of high curvature on a contour is a modification of an algorithm originally proposed by Johnston and Rosenfeld [15]. This algorithm analyzes the boundary contour's curvature at different scales, choosing local maxima along the contour. The knot points are then chosen by maximum curvature and distance along the boundary contour in order to preserve equal parametric length for opposite boundary curves.

The tangent vectors in each of the parametric directions must also be calculated. The contour of the region contains a series of three dimensional data points obtained from stereo matching that define four boundary curves on the surface. These curves are approximated by a least square cubic polynomial parametrized by arc length which is then differentiated and scaled to yield tangent vector values for the knots.

The twist vectors are more difficult to estimate. If the parametric directions on the surface are along the lines of curvature of the surface, then there is no twist in the surface and the twist vectors are zero. In practice, if care is taken, these vectors can be set to zero with minor effects on the surface. This assumes that the parametrization of the surface has been chosen wisely, with corner knot points chosen at places of high curvature or discontinuity along the boundary and spaced uniformly in both parametric directions.

6. BUILDING HIGHER LEVEL SURFACES

A level 0 surface is built from vision data only and is not an accurate description of the underlying surface. There are an infinite number of surfaces that can fit the boundary contour that vision supplies. Further, the tangents which are estimated from stereo match points are inaccurate along contours that are horizontal due to the lack of stereo match points. What is needed is information in the interior of the region to supplement the boundary information. Figure 3 describes the method of building higher level surfaces. A level 1 surface is formed by adding a tactile trace across the single surface patch defined in level 0, and a level 2 surface is formed by adding tactile traces to each of the 4 patches defined by level 1 creating a new surface with 16 patches. This method is hierarchical and general, allowing surfaces of arbitrary level to be computed. The only restriction is that the new composite surface is globally computed.

To create a level 1 surface from a level 0 surface, new knots and tactile traces across the surface must be added. A level 0 surface has a 2 x 2 knot set and 1 patch, and a level 1 surface has a 3 x 3 knot set and 4 patches. The traces begin at the point of surface contact found in the initial exploration of the region found from vision processing. The sensor then traces in the direction of the midpoints of the level 0 boundary curves. The traces preserve the equal parametrization on the surface by using the knot points at the boundary curve ends to calculate the movement direction on the surface. The points reported during these traces are combined into cubic least square polynomial curves that are differentiated and scaled to calculate the tangential information needed at the boundaries. The boundary curves tangents computed from vision data are updated to include the new tactile information, which fills in areas that lack horizontal detail from the stereo process.

Figure 4 shows the level one surface that results from active tactile sensing of the front region of the pitcher in figure 2. Figure 5 shows the same for a cereal bowl and figure 6 is for a coffee mug. The surfaces are accurate and built from sparse amounts of data. It is important to note that the vision processes are supplying the justification for building smooth curvature continuous surfaces from a region. If the region were not a smooth surface, then zero-crossings would have appeared inside the region, precluding the assumption of smoothness. The lack of zero-crossings, or the "no news is good news" criteria established by Grimson [5] supports this method and in fact is the reason it succeeds in interpolating the surfaces well.

These descriptions of surfaces are accurate and preserve the smooth character of the surfaces. Because of this and their analytic nature, stable and accurate symbolic descriptions based upon the surfaces Gaussian curvature can be computed, classifying these surfaces as planar, cylindrical or curved.

7. BUILDING FEATURE DESCRIPTIONS

If the region exploring algorithm determines that the region is a hole or cavity, a different tactile tracing routine is used to determine the boundary curve of the feature. The algorithm begins by moving the sensor just beyond the least square plane P_{lsq} of a region's contour points, aligned with $\mathbf{N}_{lsq}$. It then proceeds to move in a direction perpendicular to $\mathbf{N}_{lsq}$ until it contacts a surface. Once the surface is contacted, the sensor moves along the bounding surface staying perpendicular to $\mathbf{N}_{lsq}$, recording the contact points until it reaches the starting point of the trace.

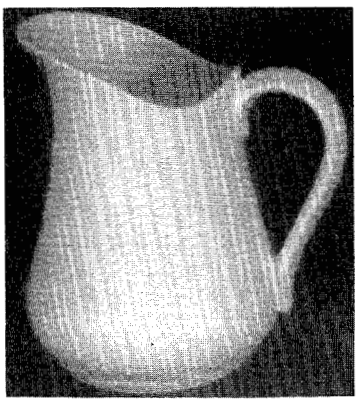
This can be a noisy procedure as many of the tactile sensor's contacts become activated in a small tight area such as the hole in the handle of a coffee mug. The spatial resolution of the sensor contacts also contributes to this phenomena. The data is not continuous, but is a set of ordered contact points that need to be smoothed and this is done by approximating the series of linked contour points with a periodic spline curve which matches derivatives at the endpoints. Figure 7 shows the smoothed boundary curve created from sensing the hole in the coffee mug. This boundary can then be used to compute cross sectional area and moments for matching against a model data base of objects.

8. SUMMARY

Discovering the three dimensional structure of an objects is important for variety of robot tasks. Single sensor systems such as machine vision systems cannot reliably compute three dimensional structure in unconstrained environments. Active, exploratory tactile sensing can be used to complement the passive stereo vision data to derive robust surface and feature descriptions of objects. The control for tactile sensing is provided by the vision system which provides regions of interest that the tactile system can explore. The descriptions of surfaces and features are accurate and have been used in a later matching phase against a model data base of objects to identify the object and its position and orientation in space.

References

1. Agin, G., "Representation and description of curved objects," Stanford University A.I. Memo, October 1972.
2. Allen, Peter, "Object recognition using vision and touch," Ph.D. Dissertation, University of Pennsylvania, September 1985.
3. Barnard, Stephen and Martin Fischler, "Computational stereo," *ACM Computing Surveys*, vol. 14, no. 4, pp. 553-572, December 1982.
4. Faux, I. D. and M. J. Pratt, *Computational geometry for design and manufacture*, John Wiley, New York, 1979.
5. Grimson, W. E. L., *From images to surfaces: A computational study of the human early visual system*, MIT Press, Cambridge, 1981.
6. Hillis, W. D., "A high resolution imaging touch sensor," *Int. Journal of Robotics Research*, vol. 1, no. 2, pp. 33-44, Summer 1982.
7. Horn, B. K. P., R. Woodham, and W. M. Silver, "Determining shape and reflectance using multiple images," *AI memo 490*, MIT AI Laboratory, Cambridge, 1978.
8. Kinoshita, G., S. Aida, and M. Mori, "A pattern classification by dynamic tactile sense information processing," *Pattern Recognition*, vol. 7, pp. 243-250, 1975.
9. Kuan, Darwin and Robert Drazovich, "Intelligent interpretation of 3-D imagery," *AI&DS technical report*, vol. 1027-1, AI&DS, Mountain View, 1983.
10. Marr, David, *Vision*, W. Freeman, San Francisco, 1982.
11. Marr, David and Ellen Hildreth, "Theory of edge detection," *Proc. Royal Society of London Bulletin*, vol. 204, pp. 301-328, 1979.
12. Nitzan, D., "Assessment of robotic sensors," *Proc. 1st International Conference on Robot Vision and Sensory Controls*, Stratford-upon-Avon, UK, April 1-3, 1981.
13. Overton, K. J., "The acquisition, processing and use of tactile sensor data in robot control," Ph.D. Dissertation, University of Massachusetts, Amherst, May 1984.
14. Ozaki, H., S. Waku, A. Mohri, and M. Takata, "Pattern recognition of a grasped object by unit vector distribution," *IEEE trans. n Systems, Man and Cybernetics*, vol. SMC-12, no. 3, pp. 315-324, May/June 1982.
15. Rosenfeld, A. and Emily Johnston, "Angle detection on digital curves," *IEEE Transactions on Computers*, vol. C-22, pp. 875-878, 1973.
16. Shneier, M., S. Nagalia, J. Albus, and R. Haar, "Visual feedback for Robot Control," *IEEE Workshop on Industrial Applications of Industrial Vision*, pp. 232-236, May 1982.
17. Tomita, Fumiaki and Takeo Kanade, "A 3D vision system: Generating and matching shape descriptions in range images," *IEEE conference on Artificial Intelligence Applications*, pp. 186-191, Denver, December 5-7, 1984.



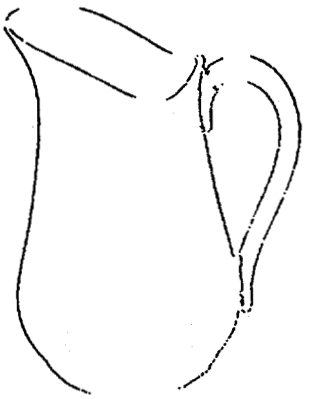


Figure 2. Pitcher and stereo match points.

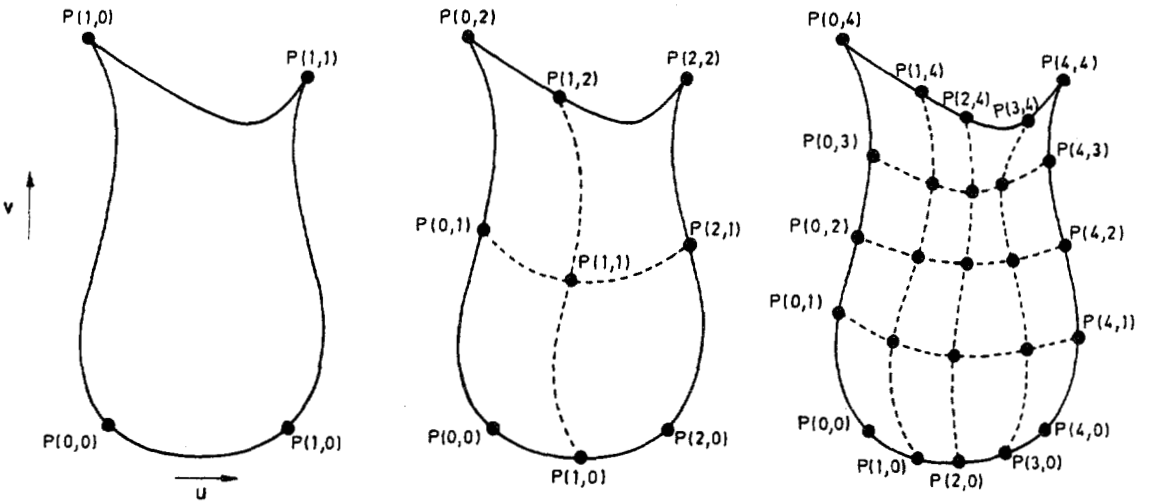


Figure 3. Level 0, level 1 and level 2 computed surfaces.

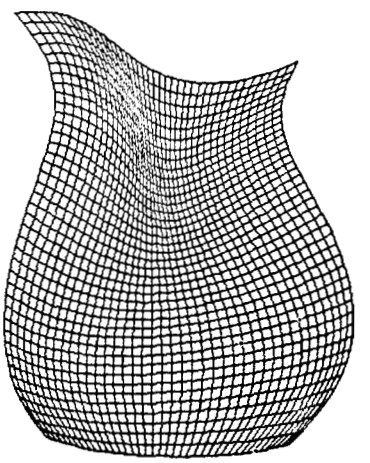


Figure 4. Computed level 1 surface for the pitcher.

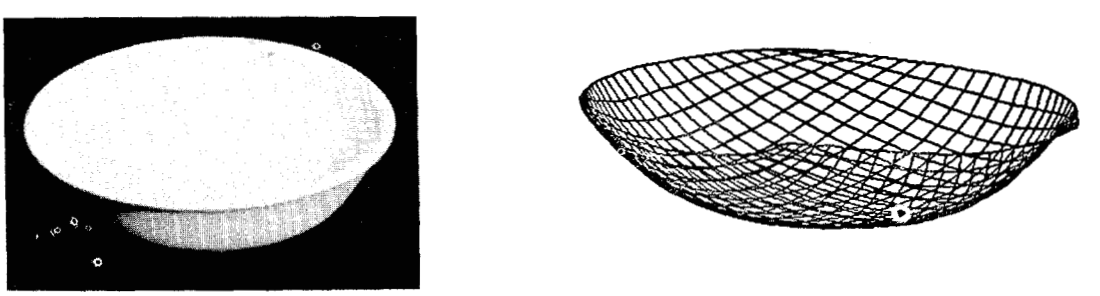


Figure 5. Cereal bowl and computed level 1 surface.

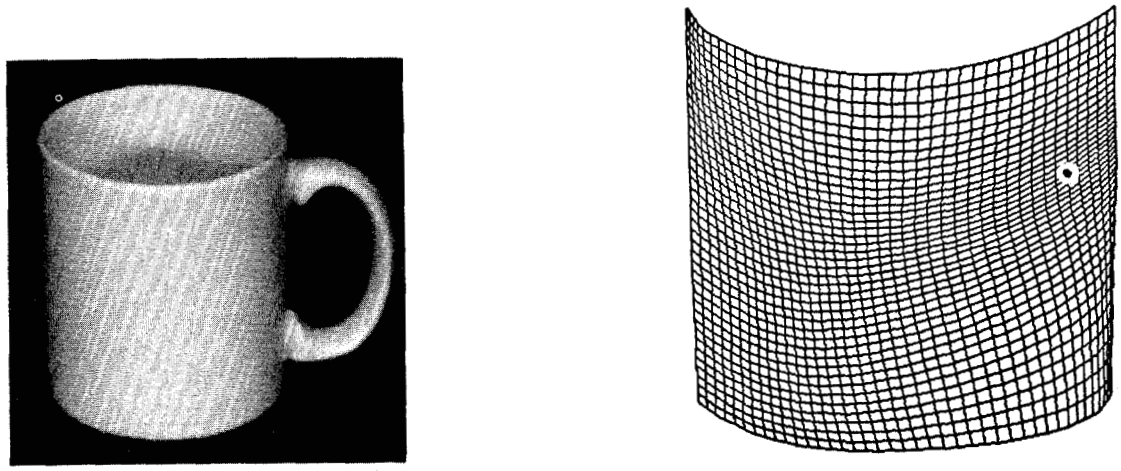


Figure 6. Coffee mug and computed level 1 surface.

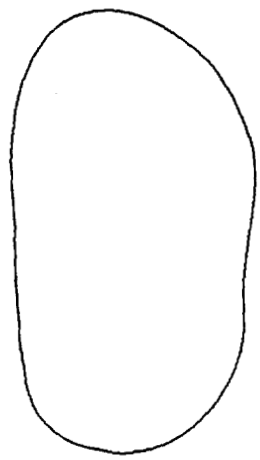


Figure 7. Traced boundary contour of the coffee mug handle hole.