Multi-scale object representation using surface patches*

Wassim Alami†‡ and Gregory Dudek†

† Centre for Intelligent Machines, McGill University
3480 University St., Montréal, Québec, Canada H3A 2A7
email: {dudek,alami}@cim.mcgill.edu

‡ currently with: CAE Electronics Ltd.
Box: 1800, St-Laurent, Québec, Canada H4L 4X4

ABSTRACT

We introduce an approach to the representation of curved or polyhedral 3-D objects and apply this representation to pose estimation. The representation is based on surface patches with uniform curvature properties extracted at multiple scales. These patches are computed using multiple alternative decompositions of the surface based on the signs of the mean and Gaussian curvatures. Initial coarse decompositions are subsequently refined using a curvature compatibility scheme to rectify the effect of noise and quantization errors.

The extraction of simple uniform curvature features is limited by the fact that the optimal scale of processing for a single object is very difficult to determine. As a solution we propose the segmentation of range data into patches at multiple scales.

A hierarchical ranking of these patches is then used to describe individual objects based on geometric information. These geometric descriptors are ranked according to several criteria expressing their estimated stability and utility.

The applicability of the resulting multi-scale description is demonstrated by estimating the pose of a 3-D object. Pose estimation is cast as an optimal matching problem. The geometric pose transformation is computed by matching two representations, which amounts to finding the three-patch correspondence that produces the best global consistency.

Examples of the multi-scale representation applied to both real and simulated range data are presented. Effective pose estimation is demonstrated and the algorithm's behaviour in the presence of noise is validated.

Keywords: scale, curvature, object representation, object recognition, surface patches, pose estimation, range imaging.

1 INTRODUCTION

For many years a key theme of computational vision has been the attempt to recover depth data from scenes. This has become a practical task in many domains using either active or passive sensing. The semantic interpretation of such data, however, remains a challenging problem. Two complementary general-purpose approaches to the description of 3-D objects are the use of either global or local parametric models.

In this paper, we will discuss the use of local parametric fitting to describe range images. It has been shown that such "patch models" based on curvature have a variety of attractive properties [12, 3, 20, 14, 2]. The segmentation of objects using patches with simple curvature properties, however, is complicated by

*This paper will appear in the proceedings of SPIE - The International Society for Optical Engineering, Boston '94.

the fact that not only is the optimal scale of processing for a single object difficult to determine, but it may not even be well defined [6]. Thus, we propose performing the segmentation at multiple scales.

The use of local parametric fitting to solve the related problems of segmentation, reconstruction, and modeling of image data has been well established. In this context and to perform image reconstruction or noise reduction from observed data, Haralick and Watson applied *facet modeling* which is based on locally fitting sub-regions of the image [7]. Besl and Jain, developed an approach to surface reconstruction by segmenting the surface into connected regions that form the seed for a region growing algorithm based on variable-order bivariate polynomials [2]. The practicality of this approach in 3-D object representation is limited by the difficulty in selecting the appropriate window of processing for a specific object and the difficulty in using higher degree surfaces (third and fourth orders) in high level computational vision tasks such as object recognition and localization. In a related approach, Boulanger et al. introduced a segmentation algorithm based on a hierarchical grouping of simple geometrical primitives into more complex ones [5]. This hierarchical structure was termed a *geometric(al) scale space* where the scale parameter is controlled by the final approximation error. The same idea of progressively increasing the descriptive complexity was common to Besl's variable order polynomial fitting [2] and is, in principle, related to minimal length encoding.

An alternative to facet modeling that is especially well suited to sparse or noisy data is based on Tikhonov regularization. Terzopoulos used *regularization* to solve the inverse problem of visible-surface reconstruction [19]. This consists of solving a variational problem that is the minimization of the energy of a *physical model.* These models are piecewise continuous, i.e. they can break occasionally to allow for discontinuities. A set of data (e.g. an object surface) is best represented in this scheme with a small number of pieces that represent the lowest overall energy. Similar energy minimization process was used by Dudek who used a curvature-tuned-smoothing (CTS) technique to interpolate and smooth the surface data in the process of describing the surface by patches having different canonical two-dimensional structure at multiple scales: convex or concave spheroidal, cylindrical, hyperbolic, etc. [6]. One drawback of such regularization-based methods is their computational cost.

The results reported here differ from those mentioned above in several respects. The most important is the fact that the objective is not to recover or reconstruct the original scene data. Our representation is explicitly focussed on reducing the image data to stable descriptors that can be effectively used in localization and recognition, perhaps with substantial information loss. Further, the segmentation we obtain through local parametric fitting and region growing at any single scale need not be completely correct. Nevertheless, by virtue of the the multi-scale nature of the representation, regions that are poorly described at one scale are typically more easily described at another (as observed by Besl, Boulanger, Witkin and many others). The surface descriptors we extract are selected to have simple curvature properties since a substantial literature suggests that curvature properties are closely associated with the necessary cues for many types of generic object recognition [1, 11, 13, 8, 15, 21, 10].

The key aspects of our approach are:

- initial surface segmentation at multiple scales,
- refinement of the segmentation into regions with uniform curvature properties,
- property estimation and quality estimation of the resulting segments, ranking of the segments across all scales.

The extraction algorithm that we propose for object representation consists of a number of steps: An initial coarse surface estimation based on a KH-mapping[1] guarantees an initial segmentation. The

[1]KH-mapping is finding the surface shape at image samples based on the signs of the mean and Gaussian curvature.

resulting segments constitute seed regions for the fitting of second order bivariate polynomials. These seed regions are grown based on a curvature compatibility scheme. The segmentation is performed at multiple scales by subsampling and processing the surface in a way to reveal its multi-scale aspect [6]. Patches at multiple scales are selected and ranked according to significance criteria, and their properties are then calculated. A set of best patches is used as an object representation. A block diagram of this process is shown in figure 1.

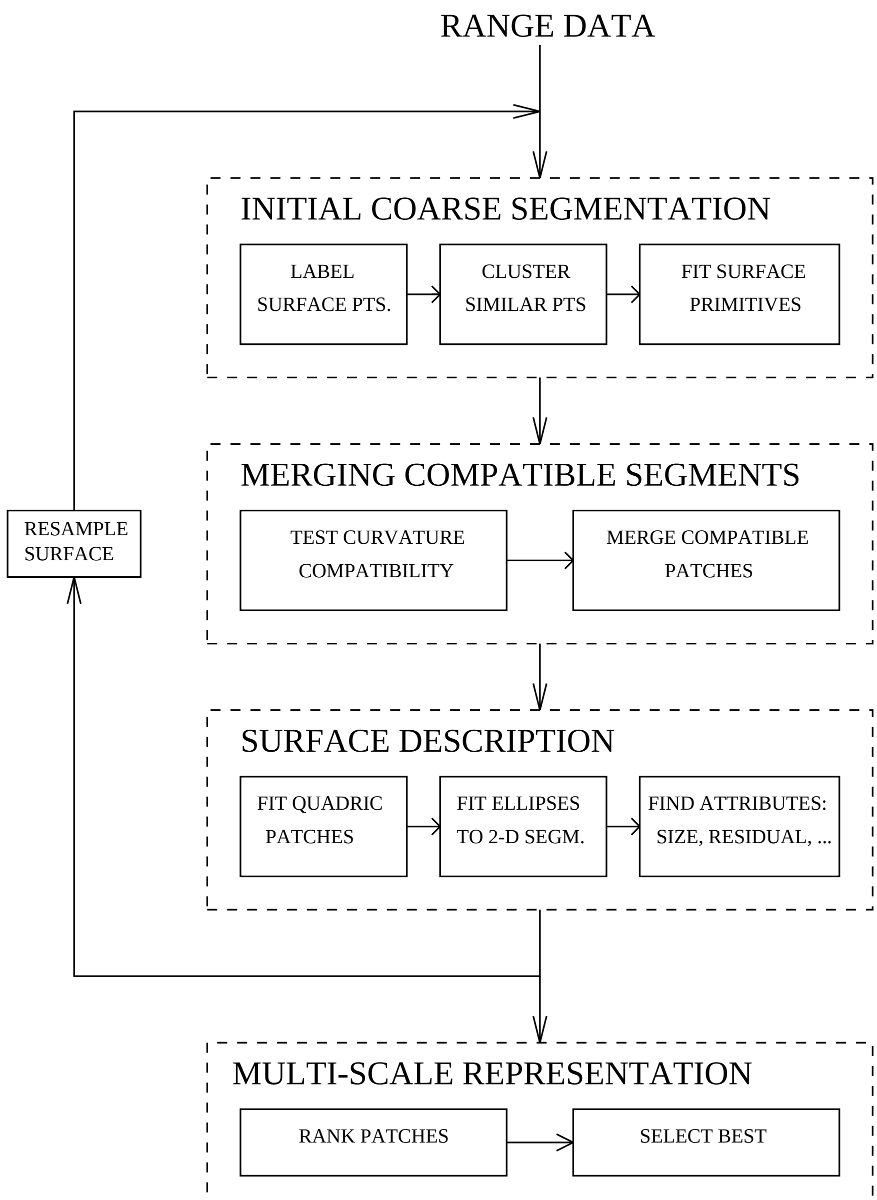


Figure 1: Stages of multi-scale object representation.

The object representation can be used to describe complex objects (not only polyhedral objects and simple machine parts). Furthermore, effective matching can be performed by selecting critical points off the object surface (to reduce the complexity and the number of features). Only these critical points, that consist of the centroids of the most stable patches, serve as the basis of a point-based matching algorithm.

2 DECOMPOSITION AND PROPERTY CALCULATION

2.1 Surface curvature

For object localization and recognition purposes, we seek features that are invariant with respect to rigid transformation and associated with perceptually relevant object properties. For these reasons, we use surface curvature as the basis for our representation.

The maximum and minimum normal curvatures at a point define the *principal curvatures*, κ_{max} and κ_{min}. The *Gaussian curvature*, K, at a point is defined as the product $\kappa_{max}\kappa_{min}$, whereas the *mean curvature*, H, is $(\kappa_{max} + \kappa_{min})/2$. Surface mean and Gaussian curvatures are invariant to rigid motion and re-parameterization.

A range image [16], considered a sampled graph surface, can be parameterized as follows: $\vec{x}(u,v) = [u\ v\ f(u,v)]^T$. Through mathematical manipulation, the mean and Gaussian curvatures can be written in the following forms:

$$K = \frac{f_{uu}f_{vv} - f_{uv}^2}{(1 + f_u^2 + f_v^u)^2}, \tag{1}$$

$$H = \frac{(1 + f_v^2) + (1 + f_u^2)f_{vv} - 2f_u - 2f_u f_v f_{uv}}{2(\sqrt{1 + f_u^2 + f_v^2})^3}, \tag{2}$$

where the subscripts designate first and second order partial derivatives. In practice, these derivatives are computed using discrete local central-difference approximations. This formulation allows the curvatures to be computed directly from range data from our sensor.

Mean and Gaussian curvatures are the basis for the segmentation of the surface into patches of constant curvature (i.e. patches defined such that the sign of H and K is constant for all adjacent pixels in a patch). Equations 1 and 2 are used to calculate H and K at every point on the surface. For noisy data, comparing measurements exactly to zero is inappropriate. The signum function

$$sgn_\epsilon(x) = \begin{cases} +1 & \text{if } x > \epsilon \\ 0 & \text{if } |x| \leq \epsilon \\ -1 & \text{if } x < -\epsilon \end{cases} \tag{3}$$

is used to compute the surface curvature sign images $sgn_\epsilon(H(i,j))$ and $sgn_\epsilon(K(i,j))$ using a preselected threshold ϵ, where ϵ is one standard deviation above the mean noise level. This pair of "images" which defines a preliminary segmentation is called the KH-map.

A label or surface type is then associated with every pixel in the image based on the signs of the mean and Gaussian curvatures according to the following equation:

$$T(i,j) = 1 + 3(1 + sgn_{\epsilon_H}(H(i,j))) + (1 - sgn_{\epsilon_K}(k(i,j))). \tag{4}$$

These labels are associated with the basic fundamental shapes which are eight in total (planar, convex or concave spherical, cylindrical, hyperbolic, etc.).

Coarse initial segmentation is obtained by finding connected regions (where the connectivity relationship is four-connectedness); that is, pixels with similar pixels adjacent to them. The *coherence* property of piecewise smooth surfaces makes pixels with the same label cluster together at the appropriate scale.

2.2 Surface description

After the range data has been segmented into regions with uniform curvature (KH-map), a refinement to the initial coarse segmentation can be performed by merging consistent segments. The resulting regions of constant curvature are fit with approximating surface functions to validate the goodness of the segmentation and to scale down the data needed to represent the surface. Hereafter, surface regions and surface fits will be referred to as *patches*.

We use biquadratic polynomials as approximating functions. These polynomials defined as:

$$f_l(x,y) = \sum_{i=0}^{m}\sum_{j=0}^{n} a_{ij}x^i y^j \tag{5}$$

where $i+j \leq 2$, are representable by a small amount of data. They are well defined over arbitrary regions in the image, and are useful for extrapolation into neighbouring regions for region growing purposes.

2.2.1 Growing Seed Patches

Curvature estimation is sensitive to noise and quantization error because it involves second order derivatives (refer to equations 1 and 2 used in calculating patch curvature). Therefore, the raw curvature estimates must be refined to be rendered meaningful and a global curvature compatibility scheme is used to do so. This is accomplished by merging and re-fitting segments determined by coarse curvature estimation when their curvatures and positions match. Surface approximating functions for seed patches are extrapolated, and the patches with the most compatible shapes are merged. In addition, a standard noise removal pass is made to remove isolated pixels assumed to be artifacts due to noise or quantization errors.

The largest patches from the initial segmentation (those larger than a threshold size α) serve as seed regions for the merging process. Quadric surfaces represented by equation 5 are fit to each seed patch obtained by initial coarse estimation and the fit error is calculated. The algorithm for merging compatible segments is explained below:

- Regions of significant size (e.g larger than α pixels) are selected for growing, labeled ω_b.
- The quantity

  $$(1/n_a) \sum_{(x,y)\in\omega_a} (z(x,y) - f(x,y,b_k))^2$$

  which is the average sum of the square difference between the function $f(x,y;b_k)$ representing region ω_{b_k} extrapolated to predict region ω_a while n_a is the number of points in the region ω_a, is calculated for all the selected regions ω_b. ω_a is merged to the region that produces the lowest extrapolation residual if it is lower than a threshold proportional to an estimate of the noise.

- The process is repeated for all the region ω_a.

2.3 Property Calculation

Once an object is segmented into surface patches that represent curvature primitives, the *geometric shape* of these patches must be encoded explicitly. Geometric properties consisting of moments and other attributes constitute our description of patch shape. Our objective is to coarsely approximate the main characteristics of each patch.

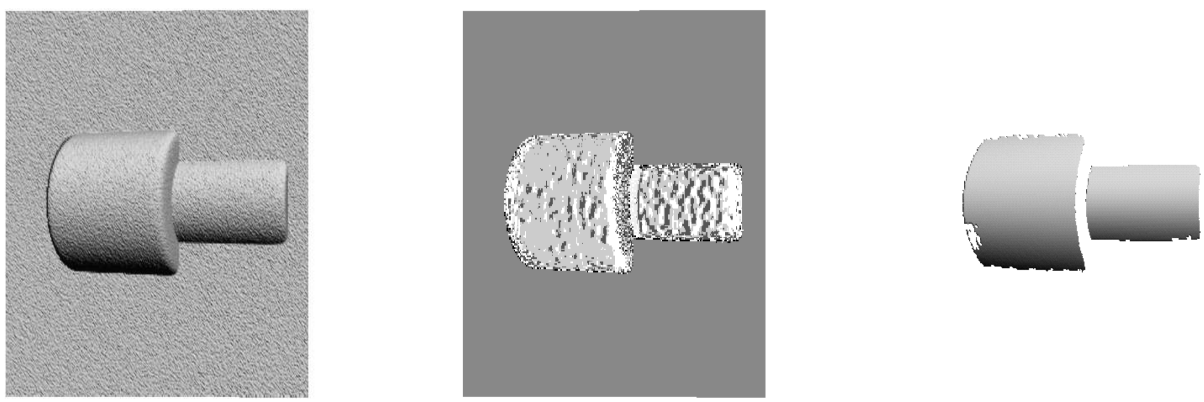
The following attributes are calculated:

- $p_{type}(S_i)$: *surface type*, there are eight basic types.
- $p_{size}(S_i)$: *size*, the area of a silhouette determined by the projection of the 3-D patch points on a plane fitting the patch.

- $p_{compactness}(S_i)$: *compactness* $(4\pi \cdot A/l^2)$, a measure of how close to a circle the silhouette is, where A is the size of the patch and l is perimeter.
- $p_{max-radius}(S_i)$: *maximum radius.*
- $p_{min-radius}(S_i)$: *minimum radius.*
- $p_{elongation}(S_i)$: *elongation* $(p_{max-radius}(S_i)/p_{min-radius}(S_i))$.
- $p_{fit}(S_i)$: *goodness-of-fit.*

Minimum and maximum radii are determined by the eigenvalues of the covariance matrix that represents the segment's points. Whereas, the goodness of fit is determined by the average residual (mean-squared error) between the polynomial approximation and the surface points.

Surface decomposition and polynomial description is applied to synthetic range data in figure 2. The robustness of surface extraction in the presence of noise is apparent in this example, where merging based on curvature compatibility has produced two stable patches from noisy initial estimates.



(a) shaded synthetic image (b) coarse image segmentation (c) surface reconstruction

Figure 2: Results for synthetic range data.

2.4 Resampling and multi-scale representation

For object representation, the surface is multiply resampled and smoothed, and the resampled surfaces are described using surface patches. A collection of patches and their attributes extracted from an object at multiple scales are denoted as the object's *multi-scale representation.*

An object's surface resampled to a uniform grid is multiply subsampled in the process of curvature calculation. Zero to multiple pixel skipping is applied in the i and j directions and the sampled pixel values are a linear average of the skipped pixels. Although the surface is subsampled in the curvature calculation process, the full surface grid is preserved for subsequent patch merging and attribute calculation.

Smoothing, which is necessary for filtering out noise and local fluctuation of the surface, has the unwanted effect of blurring the surface at discontinuities. Several investigators explored "adaptive smoothing" to eliminate the effect of blurring. In our case, we use fixed window operators and we aim at preserving the discontinuity by preserving the whole of the regions that border it.

3 MATCHING MULTI-SCALE REPRESENTATIONS

In what follows, we demonstrate the usability of multi-scale object representations by matching such representations to estimate the pose of an object in a scene.

3.1 Motion transformation

We will examine matching in terms of estimating the motion of a rigid object. In that respect, the motion parameter computation task requires the settling of four issues [17]:

1. selecting and representing the features to be used for the task,
2. representing the motion transformation,
3. establishing the correspondence between features,
4. computing the transformation in a stable and effective manner.

We select surface patches as features for matching representations. These features are represented by their geometric attributes where the centre of mass (or centroid) is considered the control point in a point-based geometric transformation. Such a transformation is represented by the equation:

$$p' = \mathbf{R}p + \mathbf{T}, \tag{6}$$

where the rotation matrix is parameterized in terms of an axis of rotation and a rotation angle about the axis. The axis and angle of rotation are found using the "adapted spherical projection" method described by Blostein and Huang [4].

3.2 An algorithm for matching

A three-point correspondence is necessary to find a geometric transformation between two three-dimensional objects. Therefore, the centroids of three patches in each representation are selected to be used as control points. The correspondence of such triplets is used for the calculation of a transformation matrix whose validity is weighted by its overall matching cost, where a low cost explains a good match. From the calculated overall costs, the pose is determined by the transformation that produces the least overall matching cost, or in other words, the best global consistency. A matching cost can be a sum of squared errors as in Mohan [9], or a sum of a special distance metric. This new distance measure involves the difference between all the properties of two patches (not simply the Eucledian distance between the two centroids) and it captures the geometric equivalence between surface patches. A precise description of the algorithm is as follows:

Select N best significant patches in each view;
For all possible sets of three correspondence pairs,
    $\mathbf{S} = \{(S_{11}, S_{21}), (S_{12}, S_{22}), (S_{13}, S_{23})\}$;
LOOP
        Find the pose $\mathcal{T}$ from the correspondence $\mathbf{S}$;
        Compute the overall matching cost based on all
            possible correspondences with respect to pose $\mathcal{T}$;
        Save the pose $\mathcal{T}$ with the smallest cost as the best pose;
END LOOP

Significant patches are those patches with a size higher than α (e.g. 50 pixels) and with a fit residual lower than some threshold ξ (ξ is proportional to the noise estimate). A set of N best patches is selected from significant patches based on the measure called *reliability* that we define below:

$$reliability(S_i) = \omega_1 p_{scale}(S_i) + \omega_2 (p_{size}(S_i)/\alpha) + \omega_3 p_{compact}(S_i) + \omega_4 p_{good-of-fit}(S_i). \quad (7)$$

The weights in this measure are set such that a "good" patch is typically a large, compact, low fit residual patch at a high scale.

4 RESULTS AND DISCUSSION

In our experiments we used range data from our movable laser range scanner (the NRC/McGill scanner); a synchronized triangulation-based scanner using a laser beam to produce a matrix of discrete depth values $z_{ij} = \tilde{g}(i, j)$ from a surface S [16]. A PUMA 560 robot arm fixed to the ceiling was used to control the position of the scanner with respect to the scene.

With these experiments, we evaluated the suitability of our representation for computing object pose. This was accomplished by either moving the range finder about the object to alter the viewpoint and hence the relative pose, or by moving the object with respect to the range finder. Multi-scale representations of the multiple views are found, and the motion is consequently calculated based on the match of the two representations at the two time instances. Below, we report the results on two distinctly different types of objects.

4.1 A collection of simple objects

A scene containing a collection of simple objects (fruits and vegetables) presents different patches that appear in the segmentation at multiple scales (refer to figure 3). To find the motion estimates for different

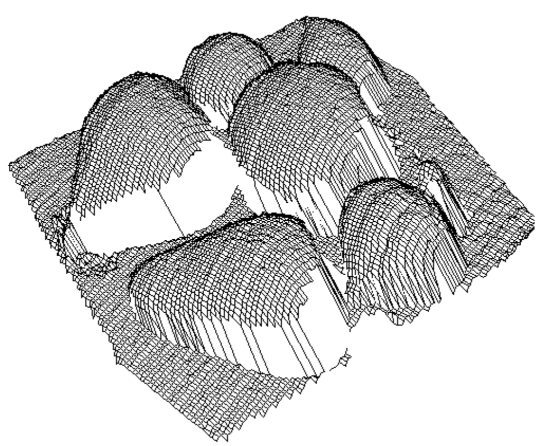


Figure 3: 3-D plot of the scene of fruits and vegetables.

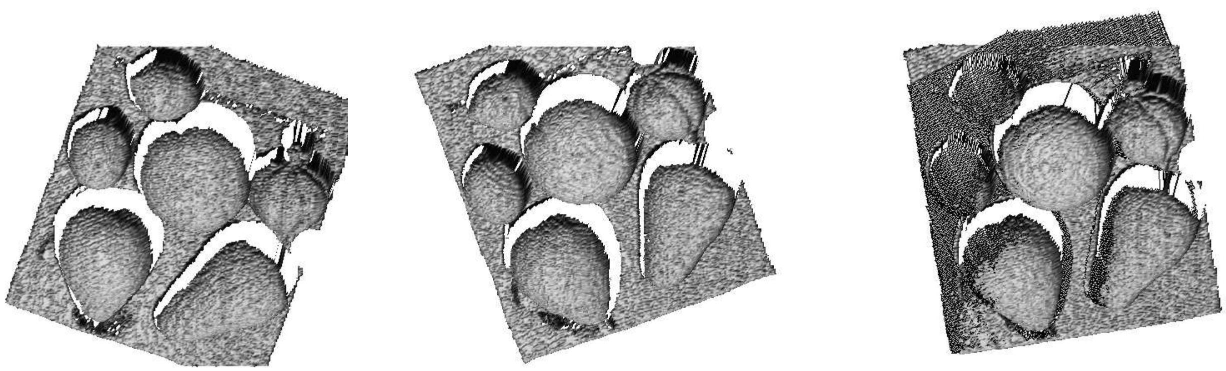
rotations and to calculate the errors involved, we perform the following set-up. The objects are fixed to the top of a rigid plate that rotates on a horizontal table. The camera is pointing perpendicular to

rotation (deg.)	10	20	30	40	50	60	70	80
error in estimate (deg.)	1.87	0.90	1.19	0.09	4.25	0.11	0.55	1.62

Table 1: Error in rotation estimates.

the table, which is assured by scanning the plate and verifying that different points on the plane are equally distant from the camera. The scene is rotated around the z axis at fixed increments. This test is performed with the camera fixed because robot motion estimates are not very accurate.

Figure 4 shows one matching case. In this figure, two distinct views are shown separate in shaded gray. Then, the two views are shown in the same coordinate system after performing the matching and the motion estimation. These views are rendered differently (black grid and shaded gray) for display purposes. The errors in the estimated rotations are shown in Table 1 for successive rotations of 10 degrees. The mean error in the estimate of orientation was 1.32 degrees.



(a) first view (b) second view (c) after matching

Figure 4: Matching experiment for a scene of simple objects. The two scenes match well despite the large motions involved.

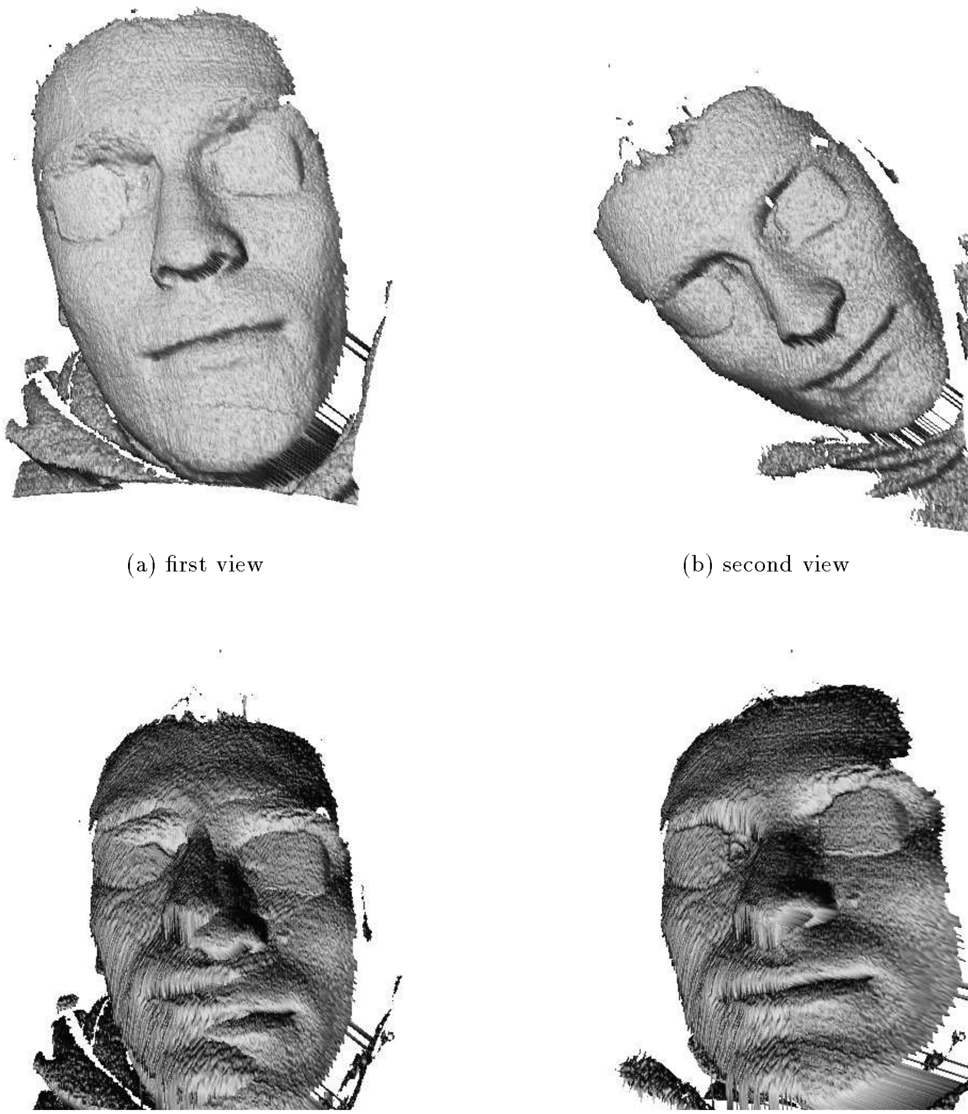
4.2 Complex object

This experiment involves the estimation of the rigid motion of a complex object such as a human face. The surface complexity evident in faces makes it very hard to extract a representationally complete set of features from a single scale. One notable characteristic of human faces in terms of segmentation and modeling is the difficulty of automatically assigning stable parts or surface descriptors at any single scale. Regions that can be extracted consistently may often be of too large a scale to provide sufficient positional constraint. Therefore, a representation that encompasses features from a wide range of scales, such as the multi-scale representation seems appropriate.

The matching of two views of the face with the camera performing a large motion between the views was performed. Key features for matching were selected automatically. These play the role of landmarks that constrain the calculation of the global geometric pose transformation. Small scale features and less stable patches serve to tune the global transformation when matched.

Although the segmentations of the two surface views are not identical, the correspondence between patches produces good results, as shown in Figure 5. Notice that the robot's motion estimate (see

Figure 5(c)) is not used as an initial estimate for the matching algorithm and is clearly much worse than our algorithm's performance. Such an estimate is used for most point-based matching algorithms to restrict the search space for point correspondences.



(a) first view

(b) second view

(c) robot's estimate

(d) algorithm's estimate

Figure 5: Fusion of the views using the robot's motion estimate as compared to using the algorithm's estimate. The results show that much better estimates are obtained through the multi-scale matching algorithm.

5 CONCLUSION

We have presented a method for the representation of 3-D objects from range measurements. We use the representation as the basis for relative viewpoint (pose) estimation. A system that successfully matches such representations to robustly estimate large-scale motions in the presence of noise has been developed.

The object representation technique is based on extracting uniform curvature surface patches and it performs the extraction at a range of scales. These multi-scale patches are obtained by alternative decompositions of the surface based on the signs of the mean and Gaussian curvatures. The initial coarse decompositions are consequently refined using a curvature compatibility scheme to rectify the effect of noise and quantization errors. Geometric attributes are then calculated for the patches. These attributes are used to rank the patches, a selection of which is used to describe individual objects.

The results obtained show that multi-scale representations can be successfully used in the localization and pose estimation of curved objects in the presence of noise. They permit pose estimation with substantial accuracy and with no prior position estimate even in cases where the data and any single scale appears unusable. This is in contrast to methods that perform accurate pose estimation from dense data but require a good *a priori* estimate [18]. As with all surface-based representations, occlusion of large parts of the object will negatively affect performance.

We have also used this representation for constrained object recognition experiments. This work is ongoing.

6 ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support of the Natural Sciences and Engineering Research Council and the Canadian Federal Centres of Excellence Program.

7 REFERENCES

[1] Fred Attneave. Some informational aspects of visual perception. *Psychological Review*, 61:183–193, 1954.

[2] Paul J. Besl and Ramesh C. Jain. Segmentation through variable-order surface fitting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(2), March 1988.

[3] Peter A. Blicher. A shape representation based on geometric topology: Bumps, gaussian curvature, and the topological zodiac. In *Proc. International Joint Conference of Artificial Intelligence*, pages 767–770, Milan, Italy, August 1987.

[4] S. Blostein and T. Huang. Estimating 3-d motion from range data. In *Proceedings 1st Conference on Artificial Intelligence Applications*. Denver, Colorado, 1984.

[5] P. Boulanger, G. Godin, M. Yamamoto, and M. Oshima. Hierarchical segmentation of range images based on optimal region growing and geometrical generalization. In *6th International Conference on Image Analysis and Processing*. Como, Italy, September 1991.

[6] Gregory Dudek and John K. Tsotsos. Shape representation and recognition from curvature. In *Proceedings of the 1991 Conference on Computer Vision and Pattern Recognition*, pages 35–41, Maui, Hawaii, June 1991. IEEE.

[7] Robert M. Haralick and Layne Watson. A facet model for image data. In *Proceedings of the IEEE Computer Society Conference on Pattern Recognition and Image Processing*. Chicago, Illinois, August 1981.

[8] D. D. Hoffman and W. A. Richards. Parts of recognition. *Cognition*, 18:65–96, 1984.

[9] N. Kehtarnavaz and S. Mohan. A framework for estimation of motion parameters from range images. *Computer Vision, Graphics, and Image Processing*, 1989.

[10] B. B. Kimia, A. Tannenbaum, and S. W. Zucker. Toward a computational theory of shape: An overview. In *Proceedings of the First European Conference on Computer Vision*, Antibes, France, May 1990.

[11] J. J. Koenderink and A. J. van Doorn. Photometric invariants related to solid shape. *Optica Acta*, 27(7):981–996, 1980.

[12] John C. Lee. Matching human faces from range data. MSc. Thesis, Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, May 1990.

[13] David G. Lowe. *Perceptual Organization and Visual Recognition.* PhD, Dept. of Computer Science, Stanford University, Sept. 1984.

[14] Ting-Chuen Pong, Linda G. Shapiro, Layne T. Watson, and Robert M. Haralick. Experiments in segmentation using a facet model region grower. *Computer Vision, Graphics and Image Processing*, 25:1–23, 1984.

[15] Whitman Richards and Donald D. Hoffman. Codon constraints on closed 2d shapes. AI memo 769, MIT AI lab, May 1984.

[16] M. Rioux, F. Blais, and J.-A. Beraldin. Laser range finder development for 3-d vision. In *Vision Interface 1989.* Toronto, Ontario, 1989.

[17] Bikash Sabata and J. K. Aggarwal. Estimation of motion from a pair of range images: A review. *CVGIP: Image Understanding*, 54(3), November 1991.

[18] G. Soucy and F.P. Ferrie. Motion and surface recovery using curvature and motion consistency. In *Proceedings of the Second European Conference on Computer Vision*, pages 222–226, Santa Marguerita Ligure, Italy, May 1992.

[19] Demetri Terzopoulos. The computation of visible-surface representations. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 10(4):417–438, July 1988.

[20] B. C. Vemuri, A. Mitiche, and J. K. Aggarwal. Curvature-based representation of objects from range data. *Image and vision computing*, 4(2):107–114, May 1986.

[21] Steven W. Zucker, Chantal David, Allan Dobbins, and Lee Iverson. The organization of curve detection: Coarse tangent fields and fine spline coverings. In *Proceedings of the 2nd Interlnational Conf. on Computer Vision*, pages 568–577, Tarpon Springs, Fla., Dec. 1988. IEEE.