

Fast 3D Object Modeling and Recognition in Range Images

Reda E. Fayek

Systems Design Engineering
University of Waterloo
Waterloo, Ontario N2L 3G1
reda@watnow.uwaterloo.ca

Andrew K. C. Wong

Systems Design Engineering
University of Waterloo
Waterloo, Ontario N2L 3G1
akcwong@watnow.uwaterloo.ca

Abstract

The recovery of three-dimensional (3D) information from range images is the major issue in the extraction and recognition of 3D objects from sensory data. 3D objects can be recognized by fitting predefined fixed models to the data and minimizing the fit error. Other techniques identify these objects by matching generic, relational and functional models to detected features in the data. Certain applications favor one approach over the other. However, the wider variety of real-world situations, which typically involve noisy data, require generic models.

Presently, it is easy to obtain edges from images, height from depth maps, but it is knowledge and constraints which are needed from a 3D scene interpretation system. This paper presents a mechanism for 3D object modeling and recognition from dense range images. We first extract symbolic information from a triangular mesh model surface approximation of the raw sensory range data. Aggregates of these features define generic object models used to abstract relevant scene information into a compact symbolic form suitable for analysis and reasoning. These resulting models are then effectively used for fast 3D object recognition. Examples of the performance of our approach are presented using both simulated and real-world data.

1 Introduction

Computer aided design (CAD) produces models which can be used for object recognition in range images. However, since the manual generation of CAD models is a laborious and expensive task, only a limited variety of complex and interesting real-world objects have CAD models. Therefore, there is a need to automatically generate CAD models of complex 3D object. Although extensive work has

been done in 3D modeling, a small fraction of it is dedicated to CAD model reverse engineering from sensory information of existing objects. This is the inverse of the CAD model design problem. Recent advances in range sensing facilitate the acquisition of dense range maps providing a wealth of 3D information. Meanwhile, the rapid growth in computing capabilities made available encourages the use of such rich data. In this paper, we present a 3D data compression mechanism to abstract the acquired sensory data into CAD models. This is a key component in object recognition systems.

We have developed a framework for 3D modeling using a hierarchy of multiresolution triangular meshes [7]. From these meshes, we detect curvature discontinuity edges and points of curvature extrema. Coarser mesh approximations are constructed by preserving these features and removing redundant sensory data. Our technique exploits these 3D topographic features to generate 3D models of controlled accuracy. This feature-based modeling technique is adequate for converting dense and noisy range data into compact symbolic information suitable for reasoning. It is particularly useful for several difficult reasoning tasks (*e.g.*, visual navigation of autonomous vehicles in outdoor terrain, replicating museum artifacts and reverse engineering of CAD models). Figure 1 gives a breakdown of the general 3D sensing, object modeling and object recognition tasks. We will illustrate the advantage of our technique for these different subtasks.

A feature in a range image of a 3D object consists of measurable properties characteristic of that object. The aggregate of the object's features forms its model. Most of feature extraction techniques are based on finding the *edges* bounding surface patches in the range images. Once these edge are found, the descriptors of their interior surface are readily measurable. This approach tends to be extremely

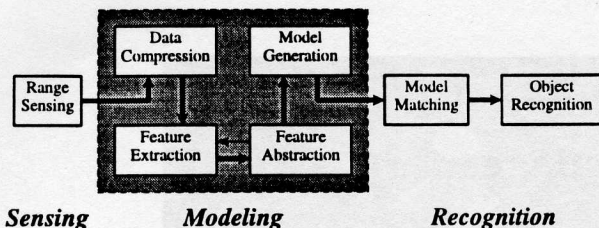


Figure 1: Sensing, modeling and recognition process and its components.

expensive when dealing with dense range images. Under these conditions (*e.g.*, a noisy range image of an indoor scene or a textured region in an outdoor scene), it may be more efficient to proceed in the opposite way: first use a *region growing* routine to detect a feature region and then extract its bounding edges. Edges in range images are classified into two basic types: *jump edges*, where the range is discontinuous, and *convex edges* or *concave edges*, where the range gradient is discontinuous.

Feature-based 3D object modeling [14] suffers from complexities due to the size and errors of real-world sensory data. Traditionally, 3D objects were often modeled by polyhedra. These are easy to analyze, are basic components of many man-made objects and can be used to approximate curved or irregular objects. Nonetheless, range images do not provide information on the internal sections of the objects; *i.e.*, they are only $2\frac{1}{2}D$ in nature. The new contribution of this work is the application of unstructured triangular meshes in synthesizing compact and generic feature-based object models suitable for reasoning applications. Most triangular mesh applications for solid modeling [9, 17, 21] were limited to visualization and rendering.

The information provided by range sensors is typically not acquired by direct measurements, but is rather computed by stereo-matching of intensity images. This introduces high noise levels. Therefore, many researchers [8, 11, 13, 19, 20] opt to use intensity information to interpret such scenes. These techniques attempt to solve two difficult problems simultaneously; (i) the mathematical projection models of 3D structures onto the captured 2D images and (ii) the interpretation of the 3D scene. When range information is available, an image interpretation system based on 3D information is far less complex. Many surface modeling techniques produce 3D surfaces from spatial data [1, 9]. While some systems interpret range sensory data directly [5, 12, 15], they are limited to clean surfaces of controlled indoor scenes. Our 3D surface model, however, lends itself for reasoning about complex,

noisy 3D scenes. Its strength stems from the efficient compression of raw range data into triangular meshes and their intelligent conversion into topographic surface features suitable for generic model matching. We will illustrate the advantage of our approach in automating the four basic modeling subtasks of Figure 1, namely, (i) data compression, (ii) feature extraction, (iii) feature abstraction and (iv) model generation. Finally, we demonstrate its experimental application to simulated and real sensory range data and demonstrate its functionality in recognizing man-made structures from aerial satellite images.

2 3D Range Data Compression by Hierarchical Triangular Meshes

The large amounts of data available from practical range sensors make it infeasible to use uniform sampling for large regions of the whole images. Therefore, a feature-based model is of essence. The variability of certain application-dependent *relevance measures*, or *discriminant properties*, in the sensory data can be used to identify important scene features and foci of interest. Our model relies on the efficient representation of such 3D surface features. This allows the segmentation of the data into *high-interest* and *low-interest* regions depending on the spatial distribution of the 3D features in the scene. High data compression ratios can be achieved by a triangular mesh representation of the data. This is due to the fact that a 3D surface described by 3D points can be approximated by a smaller number of surface elements. Computationally extensive surface manipulations become more feasible through the triangular representation.

We adopt a 3D topographic model based on the notion of unstructured, irregular triangular 3D meshes [10]. These meshes are sets of 3D points, not necessarily regularly spaced, with piecewise linear patches (*i.e.*, planar segments) approximating the actual 3D surface. The data structures describing the triangles contain attributes about the available intensity, and possibly chromatic, information. Although triangular meshes have been used to model complex 3D scenes [2], they have not been exploited to support symbolic reasoning and recognition operations. Various well-known triangular mesh generation criteria [3, 4, 16, 18] optimize different mesh properties. Unlike these methods, we construct our triangular meshes using topographic surface features [6].

Surface topographic features are automatically extracted from the meshes to form coarser triangular meshes. The iterative application of this topographic coarsening at desired resolution levels repeatedly reduces the number of points and edges not contributing to scene features. The incremental compression ratio can be subjectively controlled. The process runs until all remaining mesh elements belong to scene features, thus, achieving maximum intelligible compression. The final resulting mesh resembles a wire-frame model of the scene. Due to the irregular distribution of interesting features in a typical range data set, our irregular triangular meshes provide the desired non-uniform sampling across the covered scene. It yields lower points density in regions with little changes and higher points density in high relief areas. This model is, therefore, particularly compact, while, allowing scene analysis at different granularity levels [7].

3 Automated Feature Extraction and Abstraction

A brute force approach to feature extraction from range images would simply apply a certain operator over the whole image, thus, identifying all occurrences of a particular feature. Together with the computational cost of the subsequent analysis of the resulting features, this approach suffers from several problems related to fine tuning of the feature detection process (*e.g.*, thresholds, singularities and lack of context information). Coupling the feature extraction and feature interpretation efficiently (as in Figure 1) yields substantially better results [14]. One such approach uses several detected features to find large subsets of mutually consistent features.

While our feature-based irregular triangular mesh generation is data-driven, their interpretation is context-dependent. At the highest resolution level, we use all the raw data to approximate the perceived surface of the 3D scene. A mesh coarsening algorithm extracts primitive topographic features and preserves them in subsequent lower-resolution, compressed meshes. The surface features are topographically invariant points and edge segments in the triangular mesh approximation (*e.g.*, peak and pit points, ridge and ravine edges). Neighboring features are grouped and analyzed. Simple relations such as connectivity and continuity are studied in the local neighborhood of each feature and used to support the extraction of adjacent ones. Small isolated features usually contribute negligible global information, and overlooking them would not sub-

stantially affect the subsequent object recognition process. The detected surface features determine the induced symbolic scene features in the coarser meshes, hence the generality of this model. The interesting scene features are context-dependent and have semantic content; that is, they are subjectively defined for certain applications.

The resulting hierarchy of triangular meshes, each having a coarser representation of the scene, can be used to segment more abstract scene features (*e.g.*, flat surface patches, inclined patches, vertical patches, etc.) from the background at desired resolutions. This is performed by incrementally growing nearly-planar regions in the triangular meshes into patches with consistent quantitative attributes. These, in turn, are grouped into more complex, application-dependent symbolic primitives suitable for analytical reasoning and object recognition. This grouping is both a requirement and a merit of this model over other 3D models. We generate these nearly-planar patches by a local, incremental region-growing algorithm to limit the computational complexity.

Each unprocessed triangle in the mesh is considered as a planar patch seed. By examining its edges (and its adjacent triangles), we incrementally expand it radially outwards. If these triangles meet one of several stopping conditions, their shared patch boundary edges are "frozen" and expansion resumes elsewhere. A patch ceases to expand at a certain edge if its corresponding triangle makes a large dihedral angle with its average plane or if the edge has already been marked as a topographic feature edge. After the expansion of an edge, the patch is augmented with the corresponding exterior triangle and its average plane is accordingly adjusted. Figure 2 shows the initial and final states of such a patch. For each triangle t_k , the boundary π of the patch P_i is minimally set to the edges of t_k . The edges of π are ordered such that $\pi = \{u_i \mid u_i u_{i+1} \in \pi, 1 \leq i \leq n, u_{n+1} = u_1\}$. Then, we incrementally expand π until it reaches its maximal state. If the edge to expand $u_i u_{i+1}$ of the patch P_i is a feature edge, it is included in the maximal enclosing cycle and is not expanded further. Otherwise, if the unprocessed triangle t_k exterior to that edge meets the co-planarity criterion, it is added to P_i . When P_i cannot be expanded any further, its boundary edges are marked as feature edges and a new unprocessed seed triangle t_k is used to grow a new patch. This process continues until all triangles are grouped into contiguous patches. Ultimately, the scene is totally tessellated with nearly-planar patches.

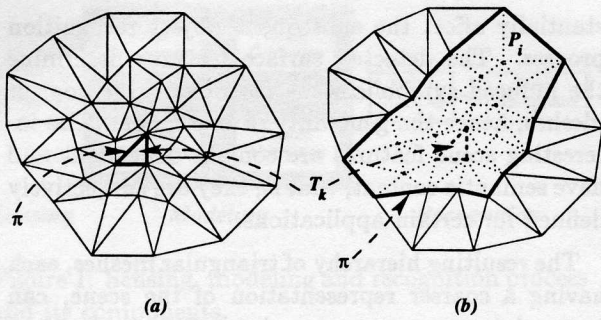


Figure 2: Growing a triangle t_k into a nearly-planar patch π : (a) initial patch, (b) final patch.

The planar patch generation exploits the detected surface features and abstracts them into more symbolic scene features. This synthesis of simple local features into global ones facilitates model building for object recognition. It effectively bridges the gap between sensory data and symbolic models. The resulting patches and their relations are the building blocks of generic 3D object models.

4 Object Model Generation

The general situation in 3D object recognition involves the attempt to match an existing object model with certain subsets of the acquired data. This problem heavily depends on the data format and on the expressiveness and complexity of the model to be matched. In order to reduce the required computational effort, the data can be abstracted and synthesized into symbolic entities, as described above. In this section, we concentrate on the generation of 3D object models suitable for the recognition process. A certain level of compatibility must be ensured between the model and the synthesized knowledge about the scene to facilitate the successful identification of objects. The automatic generation of CAD models, also known as reverse engineering of CAD models, is therefore a vital step towards object recognition.

The most straightforward fashion to formulate 3D object models is the use of wire-frames. The 3D edges in such models correspond to characteristic edges on the surface boundary of the 3D object. Wire-frame models tend to be precise, and therefore, they are suitable for the recognition of simple objects in accurate sensory data. Nonetheless, sensory information is invariably corrupt with errors and noise which cannot be accounted for by simple filtering. In such cases, precise models fail to detect the required objects. A more generic dual of wire-

frame models is based on the faces of the object. These are the surface patches bounded by closed cycles of wire-frame edge segments. The collection of such faces form a complete 3D surface boundary model of the object. Using faces to model an object has several advantages among which we cite: (i) fewer faces than edges are usually required to describe the object, (ii) a face description captures much more information than an edge description and (iii) more generic models can be built in terms of faces than in terms of edges.

In view of the above arguments, we use our triangular mesh 3D surface model to generate 3D object models. The nearly-planar patches form the faces of the objects of interest. If a wire-frame model is preferred, the boundaries of these patches can be used instead. After the initial triangulation is obtained, and its topographic surface features identified, it is iteratively coarsened. A compact model can be then defined by the patches and edges of the final coarse mesh. It includes information on the actual faces, their boundaries and their mutual geometric and topologic relations. These relations can be symbolically expressed as rules or constraints networks. Such a generic model would be suitable for subsequent object recognition tasks.

Since models constructed in this fashion are not particularly accurate, they are generic and flexible in nature. They can be used to identify objects using low-resolution, compressed triangular meshes. Once the object of interest is identified by fitting a coarse model to its corresponding region in the data set, more details can be extracted from the data using higher-resolution meshes. This scenario is particularly useful with the typically large sets of range sensory information. It amounts to the compression of the data, searching the compressed data for object occurrences, then, locally expanding data in the region of interest to reconstruct the object more accurately. The main strength of this approach, and the key factor in making it feasible, is our topography-preserving 3D data compression.

5 Experiments with Simulated Objects

We illustrate the behavior of our automated 3D object model generation by applying it to a set of simulated range readings. We arbitrarily define a certain 3D scene with one or more objects and position the simulated range sensor at an arbitrary location. The simulator produces values corresponding to those which a range finder would provide.

Although we do not add any type of noise to the data, the discrete measurements of such a simulation reflect the real-world situation. Figure 3 shows a perspective view of such a simulated scene and the resulting 3D model. For clarity, only the wire-frame format of the model is shown. The original version of the scene consists of 961 vertices corresponding to the discrete range measurements covering a region with two polyhedral objects on a flat background. The final model of the scene consists of 31 vertices only. This is to be contrasted with the minimal number of vertices (24) which defines the simulated scene. This demonstrates that the compression of the simulated data was possible within the limit of the digitization and the sampling interval.

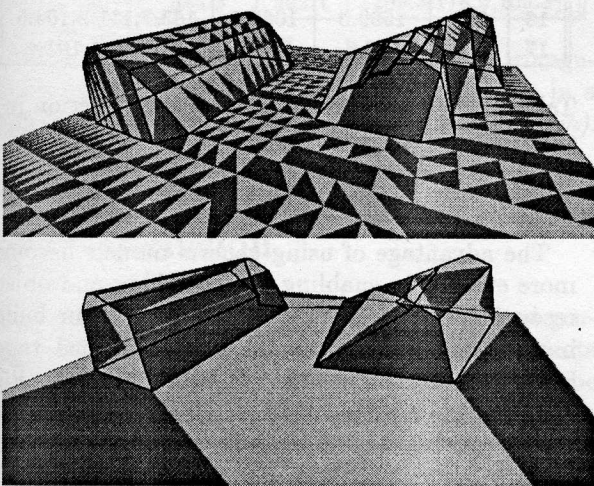


Figure 3: A simulated 3D scene of two polyhedral objects: (top) the detected feature edges of the original dense simulated range data, (bottom) the resulting wire-frame models of the objects.

6 Application to the Detection of Man-Made Structures

The versatility of our hierarchical triangular mesh model is demonstrated by its use in a wide variety of applications (*e.g.*, modeling rugged natural terrain scenes to assist autonomous vehicle navigation, automatically building CAD models of 3D objects and cartography). In this section, we demonstrate its applicability to the problem of aerial range image interpretation. This application deals with assigning meaning to certain regions of the images and segmenting its various interesting objects. Our model can be easily used both for automatic object recognition, and automatic object reconstruction of man-

made structures in real-world sensory range images.

Such unstructured environments require model-driven, seed-growing object recognition. The data is segmented into meaningful regions varying in their discriminant properties. Tentative model matching is constrained to locations where a seed region is identified. Subsequently the seed-growing phase identifies the remaining model features and recognizes any existing objects. In general, this strategy is suitable for scenes with moderate complexity, both in data and in semantics. The local nearly-planar patch generator provides the initial set of regions which can be studied quickly and reliably using semantic representations of the domain knowledge. The reliability of the intermediate-level patch representation prevents excessive backtracking [22].

In order to recognize large classes of objects, we use generic models [8, 12, 15]. We use a single generic model covering houses, apartment buildings and large hangars. When such a structure is recognized, a set of derived parameters are computed (*e.g.*, height, area of floor plan, volume, area of enclosing surface). They are used to distinguish the different objects using domain-specific knowledge and to reconstruct the detected objects' geometry. Our building model is expressed in terms of a set of nearly-planar patches and their mutual topologic and structural relations. While we use some constraints on the object's dimensions to exclude spurious matches, we allow for partial model matching. If we were to recognize outdoor vehicles using this generic model, only the thresholds of these constraints need to be changed.

We applied our technique to real geographic elevation data captured by satellite aerial imaging. The data covers an area of roughly 240×240 square meters with a 1 m resolution. The region, shown in perspective in Figure 4, features a plain with 17 buildings of similar heights. The corresponding mesh consists of 120×120 points Digital Elevation Map (DEM) regularly sampled at 2 m intervals. The detected houses' walls and roofs, shown in grey, are formed by adjacent nearly-planar patches satisfying our generic model. For each patch, in addition to its location, the process also computes its total surface, its normal vector (*i.e.*, its 3D orientation), its boundary and its neighboring patches. Together, these are used to reconstruct the geometry of the extracted objects if needed. Table 1 gives the results of these computed parameters. For each house, they include the number of patches forming it, the area of the total surface enclosing it, its floor area and its center of gravity (*C.O.G.*). All dimensions are in meters in a right-hand coordinate system.

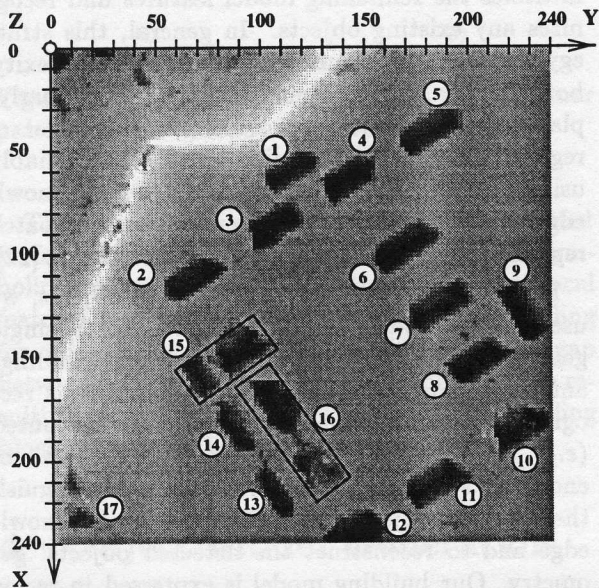


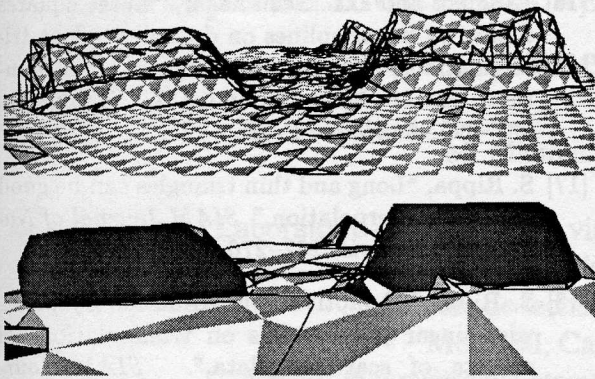
Figure 4: Extraction of houses in a suburban area [120×120 points, sampled at 2 m.]: range image data in grey-scale format (*left*), original dense triangular mesh (*higher right mesh*) and detected houses (*lower right mesh*).

#	Patches	Outside surface	Floor area	C.O.G.
1	51	618.9	409.9	60.8,110.8,103.3
2	59	825.7	512.5	111.3,64.5,104.1
3	57	818.9	511.6	86.7,103.1,104.7
4	76	835.0	514.0	65.0,137.6,104.7
5	47	860.1	535.1	42.4,176.8,105.2
6	78	929.3	596.1	97.1,164.9,105.1
7	80	932.6	634.2	125.0,180.5,105.0
8	90	895.4	590.0	150.8,198.7,105.0
9	85	821.3	534.7	128.0,220.5,105.1
10	77	915.7	604.4	183.8,217.1,105.1
11	63	885.7	588.4	209.3,179.0,105.3
12	58	631.2	436.3	229.5,141.9,105.4
13	98	609.7	400.3	212.1,101.7,103.6
14	91	617.3	415.9	181.8,83.3,103.7
15	165	1510.9	997.7	149.7,79.7,104.4
16	218	1599.0	1065.8	183.7,111.2,104.5
17	81	533.6	366.2	218.0,9.1,101.4

Table 1: Numerical results of house extraction in a suburban area.

The advantage of using generic models becomes more evident in enabling the matching and object recognition at various resolution levels. Our building 3D model can be applied to coarsened topographic meshes rather than the full data sets. Figure 5 depicts a subset of the original test region featuring two houses. The mesh consists of raw data given in a 40×40 DEM sampled at 2 m intervals. It was coarsened several times reducing the number of vertices from 1600 to 283 (over 82% saving). The numeric results of the house detection in this coarser mesh are almost identical to those obtained using the high resolution mesh in the previous example. These two houses correspond to houses 10 and 11 in Figure 4. Their shape and properties are preserved despite the mesh coarsening.

These results show that our hierarchical triangular meshes, together with the generic model of man-made structures, can successfully handle real world sensory data. Range data in general, and in remote sensing and photogrammetry in particular, is typically imperfect. Our approach reduces the 3D object recognition complexity through intelligent data compression and filtering by triangular mesh topographic coarsening. While object identification is performed with the compressed data, object reconstruction can be subsequently done at any desired resolution level. For accurate object recognition, the high resolution meshes are used. The obtained results provide good initial guesses for techniques combining range and intensity data.



#	Patches	Outside surface	Floor area	C.O.G.
1	83	915.9	603.5	(183.8,217.1,105.1)
2	62	884.3	592.1	(209.3,178.9,105.3)

Figure 5: Extraction of houses: original triangular mesh (*top mesh*) and the detected houses in a coarser mesh with 82% less vertices (*bottom mesh*).

7 Conclusion

We described a mechanism for the automated generation of triangular mesh models for 3D objects from real-world sensory data. Such models can be generally expressed in terms of the objects' corresponding wire-frames or their surface patches. Our 3D scene model provides support for symbolic analysis and interpretation of complex real-world scenes such as outdoor aerial range images. The range data is approximated by a hierarchy of irregular triangular meshes of decreasing resolutions. From these, the scene features are represented by collections of neighboring nearly-planar patches which are grouped into more complex symbolic scene entities (man-made structures in this case) suitable for analytical reasoning. The attributes of these patches are used to accurately reconstruct the identified objects. They can also be expressed at various resolution levels allowing a symbolic generic description of the objects' characteristic features.

Our feature-based model has the advantage of allowing the recognition to be performed using the compressed data, thus, substantially reducing the model matching complexity. Furthermore, significant storage savings are made possible by its 3D data compression. After the initial object identification is done, more accurate 3D object reconstruction can be performed, if required, using the uncompressed original data, or any intermediate compressed levels.

Future efforts will concentrate on extending the current system with more complex symbolic manipulations of the extracted nearly-planar patches in order to support more intricate scenes. The generic object models used thus far can be slightly modified to detect classes of similar objects. More generic models will be built to deal with substantially different scenarios. Additionally, the raw data used thus far only considers 3D range information. Various physical surface properties can also be obtained and used to enrich the models.

References

- [1] R.C. Bolles and P. Horaud, "3DPO: A three-dimensional part orientation system," *International Journal on Robotics Research*, 5(3):3-26, 1986.
- [2] M. Cazzanti, L. De Floriani, E. Puppo and G. Nagy, "Visibility computation on a triangulated terrain," *6th International Conference on Image Analysis and Processing*, 721-728, 1992.
- [3] L.P. Chew, "Guaranteed-quality mesh generation for curved surfaces," *9th Annual Symposium on Computational Geometry*, 274-280, 1993.
- [4] N. Dyn, D. Levin and S. Rippa, "Data dependent triangulations for piecewise linear interpolation," *IMA Journal of Numerical Analysis*, 137-154, 1990.
- [5] T.-J. Fan, G. Medioni and R. Nevatia, "Recognizing 3-D objects using surface descriptions," *IEEE Transactions on PAMI*, 11(11):1140-1157, 1989.
- [6] R.E. Fayek and A.K.C. Wong, "Triangular mesh model for natural terrain," *SPIE Intelligent Robots and Computer Vision XIII*, 2953, 86-95, 1994.
- [7] R.E. Fayek and A.K.C. Wong, "Hierarchical model for natural terrain using topographic triangular meshes," *IASTED International Conference on Robotics and Manufacturing*, 252-255, 1995.
- [8] P. Fua and A.J. Sander, "An optimization framework for feature extraction," *Machine Vision Applications*, 59-87, 1991.

- [9] P. Fua and P.T. Sander, "From points to surfaces," *SPIE Geometric Methods in Computer Vision*, 286-296, 1991.
- [10] R.J. Fowler and J.J. Little, "Automatic extraction of irregular digital terrain models," *Computer Graphics (SIGGRAPH'79 Proceedings)*, 13(2):199-207, 1979.
- [11] A. Heurtas and R. Nevatia, "Detecting buildings in aerial images," *Computer Vision, Graphics and Image Processing*, 41:131-152, 1988.
- [12] A.K. Jain and R. Hoffman, "Evidence-based recognition of 3-D objects," *IEEE Transactions on PAMI*, 10(6):783-802, 1988.
- [13] R. Mohan and R. Nevatia, "Using perceptual organization to extract 3-D structures," *IEEE Transactions on PAMI*, 11(11):1121-1139, 1989.
- [14] D. Nitzan, R. Bolles, J. Kremers and P. Mulgaonkar, "3-D Vision for robot applications," *Machine intelligence and knowledge engineering for robotic applications, NATO ASI Series, F33*, 21-81, 1987.
- [15] A.P. Pentland, "Automatic extraction of deformable part models," *International Journal on Computer Vision*, 4:107-126, 1990.
- [16] E. Quak and L.L. Schumaker, "Least squares fitting by linear splines on data dependent triangulations," *Curves and Surfaces*, P.-J. Laurent, A. Le Méhauté and L.L. Schumaker, 387-390, Academic Press, 1991.
- [17] S. Rippa, "Long and thin triangles can be good for linear interpolation," *SIAM Journal of Numerical Analysis*, 257-270, 1992.
- [18] S. Rippa, "Adaptive approximation by piecewise linear polynomials on triangulations of subsets of scattered data," *SIAM Journal on Scientific and Statistical Computing*, 1123-1141, 1992.
- [19] M. Roux and D.M. McKweon, "Feature matching for building extraction from multiple views," *IEEE Computer Vision and Pattern Recognition*, 46-60, 1994.
- [20] J.A. Shufelt and D.M. McKweon, "Fusion of monocular cues to detect man-made structures in aerial images," *Computer Vision, Graphics and Image Processing*, 57(3):307-330, 1993.
- [21] W.J. Schroeder, J.A. Zarge and W. E. Lorensen, "Decimation of triangle meshes," *Computer Graphics*, 26:65-70, 1992.
- [22] P. Suetens and P. Fua and A.J. Hanson, "Computational strategies for object recognition," *ACM Computing Surveys*, 24(1):5-61, 1992.