INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6” x 9” black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600
BUILDING RECOGNITION FROM AERIAL IMAGES BY USING ABDUCTION

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

the Degree Doctor of Philosophy in the Graduate

School of The Ohio State University

By

Jee-cheng Wu, M.A., M.S.

*****

The Ohio State University
1997

Dissertation Committee:

Dr. Anton F. Schenk, Adviser

Dr. Alan Saalfeld

Dr. John R. Josephson

Approved by

Adviser
Graduate Program in Geodetic Science and Surveying
ABSTRACT

Buildings are found in almost all kinds of cartographic mapping and their robust detection and delineation require techniques that exploit knowledge about man-made features. This research proposed an approach by using 3-D surface patches to solve the problem of identifying buildings from aerial images. It is based on both data driven and hypotheses driven processing to explain the surface patches. Based on the characteristics of buildings and Schenk’s model of layered abductive building recognition, a structure-based generic building model, top-down abductive reasoning, and two phases of bottom-up processes were proposed. A structure-based generic building model focuses on part/whole and spatial/geometric relationships between buildings and nearby objects. The top-down abductive reasoning is based on the unique property of aerial images, namely the top view of a scene, which implies that finding a building is to find roofs of a building. And the roofs are hypothesized with 3-D flat or curved surfaces. The two phases of bottom-up processes include (1) reconstructing surfaces from a pair of aerial images, (2) detecting humps from the reconstructed surface, and (3) computing properties of surface patches. Finally, four suggestions are made for further research: using multi-sensors data, introducing reliable weights associated with hypotheses, more clearly describing task-specific knowledge, and researching on machine vision problems.
To my parents,
ACKNOWLEDGMENTS

I wish to acknowledge the guidance and the extreme patience shown to me by my advisor, Dr. Anton F. Schenk, during my years in graduate school and on the work of this dissertation. Thanks should also be made to the other members of my committee, Dr. Alan Saalfeld and John R. Josephson for their advice on this dissertation. I am very grateful to my wife, Jui-chin Liu, for accomplishing the enormous task of correcting and polishing my English, and for the enormous emotional support offered to me in the preparation of this dissertation. Finally, to my son, Kahn-bao Wu, I express my sincere thanks for providing me with the ultimate motivation for completing this dissertation.
VITA

August 1, 1956 ............................................. Born - Keelung, Taiwan

1979 .............................................................. B. A., Chung Cheng Institute of Technology


1993-1996 .................................................... M. S., The Ohio State University

PUBLICATIONS


FIELDS OF STUDY

Major Field: Geodetic Science and Surveying
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>iv</td>
</tr>
<tr>
<td>Vita</td>
<td>v</td>
</tr>
<tr>
<td>List of Figures</td>
<td>ix</td>
</tr>
<tr>
<td>Chapters:</td>
<td></td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Background and motivation</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Statement of the Problem and Proposed Strategy</td>
<td>3</td>
</tr>
<tr>
<td>1.3 Research Contributions and Scope</td>
<td>6</td>
</tr>
<tr>
<td>1.4 A Guide to Reading this Dissertation</td>
<td>6</td>
</tr>
<tr>
<td>2. Literature review</td>
<td>8</td>
</tr>
<tr>
<td>2.1 Building Recognition from a Single Aerial Image</td>
<td>9</td>
</tr>
<tr>
<td>2.2 Building Recognition from a Pair or Multiple Aerial images</td>
<td>12</td>
</tr>
<tr>
<td>2.3 Conclusion on Building Recognition</td>
<td>15</td>
</tr>
<tr>
<td>3. The Problem Domain of Building Recognition</td>
<td>17</td>
</tr>
<tr>
<td>3.1 Visual Perception</td>
<td>18</td>
</tr>
<tr>
<td>3.1.1 Issues on the Theory of Low-Level Vision</td>
<td>21</td>
</tr>
<tr>
<td>3.1.2 High-Level Vision Processes</td>
<td>22</td>
</tr>
<tr>
<td>3.1.3 The Combination of Bottom-Up and Top-Down Processing</td>
<td>24</td>
</tr>
<tr>
<td>3.2 Object Recognition by Using Machine Vision</td>
<td>26</td>
</tr>
<tr>
<td>3.2.1 Issues in Object Recognition</td>
<td>26</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Representation</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Recognition</td>
</tr>
<tr>
<td>3.2.4</td>
<td>A Note on Object Recognition</td>
</tr>
<tr>
<td>4.</td>
<td>Knowledge Representation and Abduction</td>
</tr>
<tr>
<td>4.1</td>
<td>Knowledge</td>
</tr>
<tr>
<td>4.2</td>
<td>Knowledge Representation</td>
</tr>
<tr>
<td>4.3</td>
<td>Abduction</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Josephson's Account of Abduction</td>
</tr>
<tr>
<td>4.3.2</td>
<td>The Generation of Possible Hypotheses</td>
</tr>
<tr>
<td>4.3.3</td>
<td>The Selection of Plausible Hypotheses</td>
</tr>
<tr>
<td>4.3.4</td>
<td>The Combination of Plausible Hypotheses</td>
</tr>
<tr>
<td>4.3.5</td>
<td>A Note on Induction and Abduction</td>
</tr>
<tr>
<td>5.1</td>
<td>A Structure of the Features on the Earth's Surface</td>
</tr>
<tr>
<td>5.2</td>
<td>The General Characteristics of Buildings</td>
</tr>
<tr>
<td>5.3</td>
<td>A Proposed Generic Building Model</td>
</tr>
<tr>
<td>5.4</td>
<td>Proposed Hypotheses and Assumptions for Building Recognition</td>
</tr>
<tr>
<td>5.5</td>
<td>Proposed Reasoning and Process Strategies for Building Recognition</td>
</tr>
<tr>
<td>5.5.1</td>
<td>Schenk's Model of Layered Abductive Building Recognition</td>
</tr>
<tr>
<td>5.5.2</td>
<td>A Proposed Reasoning Strategy</td>
</tr>
<tr>
<td>5.5.2.1</td>
<td>What Kinds of Cues are Used?</td>
</tr>
<tr>
<td>5.5.2.2</td>
<td>The Use of 3-D Surfaces</td>
</tr>
<tr>
<td>5.5.2.3</td>
<td>What Hypothesis is Used?</td>
</tr>
<tr>
<td>5.5.3</td>
<td>A Proposed Process Strategy</td>
</tr>
<tr>
<td>5.5.3.1</td>
<td>Determining Levels of Process</td>
</tr>
<tr>
<td>5.5.3.2</td>
<td>Decomposing The Task</td>
</tr>
<tr>
<td>5.5.3.2.1</td>
<td>The First Phase of Process --</td>
</tr>
<tr>
<td>5.5.3.2.2</td>
<td>The Second Phase of Process --</td>
</tr>
<tr>
<td>5.5.4</td>
<td>A Note on Surface Recognition and Segmentation</td>
</tr>
<tr>
<td>6.</td>
<td>Experimental Results</td>
</tr>
<tr>
<td>6.1</td>
<td>An Example of the First Phase Process</td>
</tr>
<tr>
<td>6.2</td>
<td>Examples of the Second Phase Process</td>
</tr>
<tr>
<td>6.3</td>
<td>Discussion</td>
</tr>
<tr>
<td>7.</td>
<td>Conclusion and Future Research</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Basic Framework of Scene Recognition in Mapping</td>
</tr>
<tr>
<td>3.1</td>
<td>A Generic Visual Processing Model</td>
</tr>
<tr>
<td>3.2</td>
<td>Marr's Vision Paradigm</td>
</tr>
<tr>
<td>3.3</td>
<td>Ullman's Visual Routines</td>
</tr>
<tr>
<td>4.1</td>
<td>Josephson's Account of Abduction</td>
</tr>
<tr>
<td>5.1</td>
<td>A Simplified Classification of the Features on the Earth's Surface</td>
</tr>
<tr>
<td>5.2</td>
<td>An Instance of Semantic Network of Generic Building Model</td>
</tr>
<tr>
<td>5.3</td>
<td>The Hypotheses of an Instance of Generic Building Model in a Suburban area</td>
</tr>
<tr>
<td>5.4</td>
<td>Schenk's Model of Layered Abductive Building Recognition</td>
</tr>
<tr>
<td>5.5</td>
<td>A Reasoning Strategy for the Proposed Building Recognition</td>
</tr>
<tr>
<td>5.6</td>
<td>The First Phase of Process</td>
</tr>
<tr>
<td>5.7</td>
<td>The Second Phase of Process for Surface-based Recognition</td>
</tr>
<tr>
<td>6.1</td>
<td>The Results of the First Phase of Process</td>
</tr>
<tr>
<td>6.2</td>
<td>Segmentation of a Synthetic Mountain-roof Image</td>
</tr>
<tr>
<td>6.3</td>
<td>Segmentation of a Test Range Image</td>
</tr>
<tr>
<td>6.4</td>
<td>The Results of the Second Phase Process</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

The single most time consuming aspect of many projects on photogrammetric mapping and geographic information systems (GISs) is capturing feature data obtained from aerial photography. In photogrammetric mapping, the stereocompilation of detail in topographic maps can be started when a stereo model has been completely oriented. Photogrammetric mapping is generally divided into two principal phases: planimetric features and terrain surface. Currently, the collection of spatial features is inherently a highly manual process with thousands of human operator actions performed during the workday.

The use of analogic and analytic plotters to extract and interpret features in a stereo model depends on human operator experience. Feature extraction and interpretation are usually the most time-consuming tasks in photogrammetric mapping. The recognition of object features in a digital image, which is related to the domains of computer vision (CV) as well as image understanding (IU), is a task of increasing
importance with the trend in digital photogrammetry (DP) towards automatic mapping (Schenk, 1993). Because the automatic extraction of significant man-made features such as buildings and roads from aerial images is a complex problem, human interaction still remains essential in the DP process. However, the slow acceptance of digital photogrammetry workstations (DPWs) is a direct consequence of the lack of clear advantages over currently used analytical systems. DPWs do not provide a more cost effective option than analytic plotters, although the automatic function of DPWs is ortho-rectification, which reliably produces a successful product such as orthophotos. It is believed that DP will play an important role in the development of automation when DP is used in map production.

For the past ten years, DP researchers have been focusing on image matching to do such tasks as automatic orientation, automatic aerial triangulation, and automatic digital terrain model (DTM) generation, using many different approaches such as cross correlation, least squares matching, and object space matching. For example, combining feature-based matching with precise point determination results in a reliable and robust method for automatic orientation (Schenk et al., 1991; Heipke, 1997). Nevertheless, this emphasis seems to be changing. For example, recent developments indicate that DP should not only cover the geometric relations of an image but also the semantic aspects such as recognizing a building or a road (Schenk, 1994).

Identifying a building in a stereo model and digitizing its corners are routinely performed by human operators with great ease, leading to underestimation of the difficulty in solving them automatically. Semantic grouping involves human visual
perception and prior knowledge, but it has proven quite difficult to develop systems that perform visual tasks in a general manner. Currently, DP is in a state of great intellectual vibration and paradigm shift; it is busy absorbing many new methods and has been assuming a prominent role as an integrator of some sciences and engineering techniques. Although automatic extraction and interpretation of man-made features obtained from aerial images are far from being practical, there is encouraging development in making optimum use of multiple sources and methods to achieve scene recognition. The ultimate aim of developers must be to achieve an automatic photogrammetric system, which will remove the human operator and be cost effective, as shown in Figure 1.1.

![Diagram of scene recognition framework](image)

**Figure 1.1. Basic Framework of Scene Recognition in Mapping**

### 1.2 Statement of the Problem and Proposed Strategy

The problem addressed in this research is stated as follows:

A pair of epipolar aerial images and their DTM data are provided in order to develop a reliable procedure to interpret the image. The ultimate goal is to recognize the buildings in the image.
While the recognition of all cartographic features for fully automated machine vision is far beyond the current capability of DP technology, single feature recognition (such as building recognition) is feasible. Buildings are found in almost all kinds of cartographic mapping and their robust detection and delineation require techniques that exploit knowledge about man-made features. There are a variety of techniques to generate and verify building hypotheses, involving such knowledge as edge-line analysis, shadow analysis, stereo disparity analysis, perceptual organization, model-based matching, and the fusion technique. However, automation has not been fully successful in the extraction of buildings from images, although some researchers have suggested that automated feature extraction would ultimately be useful.

As for the simple discrimination of objects in the visual field, people may rely on the interaction of characteristics or cues such as size, shape, color, position in the field, light and shade, and stereoscopic vision. However, it is generally accepted that the human levels of performance in image understanding will require the use of massive computational resources as well as the representation and the use of massive amounts of domain knowledge. Combining the results from multiple sources (e.g. gray image, color image, range image) of possibly different resolutions relying on texture and shadow analysis, image matching, geometric constraints, has been suggested. For example, Shufelt and McKeown (1993) describe a useful cooperative method, the fusion technique, in approaching the building extraction problem. In addition, it is generally accepted that the three-dimensional terrain shape also plays a key role in the extraction of buildings (Haala, 1994; Weidner and Forstner, 1995).
The properties that enable visual perception are numerous and multifaceted, thus devising machine vision is a formidable task. Before the complex problems of machine vision can be overcome, machine vision needs to be sought in different places and at different levels of function. For example, the human visual system has been broadly divided into low-level and high-level processing. Low-level processing is based on inborn human capabilities, while high-level processing is based on such factors as previous experiences, knowledge, and social culture. It is reasonable to assume that currently no single extraction method will perfectly delineate a building in every scene because each method may only hypothesize a subset of information necessary to produce an improved interpretation of a building in a given scene.

Appropriate use of knowledge representation (KR) in a given problem can greatly simplify the problem-solving mechanisms needed to achieve machine vision. Good KR is well suited for a task such as building recognition because KR exposes the inherent constraints of a problem and facilitates the solution to the problem. The key issue of building recognition is reasoning, which intelligently allows the system to sift through the data, reduce dimensions, identify the patterns of interest, and then order the system to find other instances or similar occurrences. Thus, recognition becomes a formal process of determining the logical consistency of geometric and radiometric evidences that are derived from a number of sources and steps.

To achieve accurate building recognition, we need to focus on the concept of simplification; that is, we should devise a practical solution incorporating a general approach and existing methods. Therefore, a building recognition reasoning strategy is
evaluated by asking the following questions: (1) What kinds of knowledge are required? (2) How is knowledge represented? and (3) What criteria are used to determine the recognition accepted? While these three questions sound simple, they are difficult to answer because we don't have a comprehensive theory of visual perception to follow. In this research, the above questions are answered by using abduction. Our goal is to find the "best" set of building recognition hypotheses which will allow us to recognize different kinds of buildings on aerial images. The proposed building recognition hypotheses will be discussed in Chapter Five.

1.3 Research Contributions and Scope

The contributions and scope of this dissertation include two parts. First, how the third dimensional information (depth) is directly related to the task of building recognition is explicitly described. Second, a generic building model and a two-phase process strategy using abduction are introduced to accomplish the task of building recognition.

1.4 Organization of the Dissertation

The remainder of this dissertation is organized as follows. Chapter Two consists of a literature review on building extraction techniques, focusing on the strategies and hypotheses used. Chapter Three briefly describes the problem domain of building recognition, seeing and thinking. In considering the building recognition problem in the visual perception and model-based object recognition domain, there are two difficulties:
incomplete theories of vision and the need to model a huge number of existing buildings. Chapter Four describes the relationship between knowledge representation and abduction. Since abduction is a way of forming hypotheses to explain observation data, it allows reasoners to convert episodic information (e.g. scheme, script) instantly into semantic information. Chapter Five focuses on representing a generic building model, finding primitive features from a pair of aerial images, and generating a small set of hypotheses as well as a reasoning strategy with two-phase process to recognize buildings in a scene. In Chapter Six, a few experiments are illustrated and the examples include synthetic and real images. Finally, the overall conclusions of the research are provided in Chapter Seven, along with suggestions for improving current research, as well as recommendations for future research directions.
CHAPTER 2

LITERATURE REVIEW

Recent DP research has focused on the application of machine vision techniques to recognize buildings in aerial images. An aerial image contains a real world scene of two dimensions, in which many objects with a variety of shapes are included. In addition, the surfaces of these objects contain a variety of textures and reflectance. Furthermore, owing to trees, shadows, or other objects, some of the surfaces may be occluded. All of these factors increase the difficulty of extraction or recognition.

This chapter provides a brief overview of works related to building recognition systems devoted to acquiring either explicit or implicit concept descriptions of buildings. It is believed that knowledge in the problem domain is more highly prized than knowledge in the tool domain because tools change considerably over time, while the problem domain knowledge changes more slowly. However, problem solutions often reflect the character of the tools that are used to solve them.
2.1 Building Recognition From A Single Aerial Image

An early study by Tavakoli and Rosenfeld (1982) focuses on a method of building and road extraction from aerial photographs. Straight line segments, gray level, and geometric condition are used to constrain the search when possible, and the segments are grouped into building-like and road-like features.

Probably the best known building recognition systems were developed at the University of Southern California (USC) and Carnegie-Mellon University (CMU). At USC, Huertas and Nevatia (1988) developed rectangle hypotheses with local contour tracing techniques to detect buildings in aerial images. Shadows were used to verify the hypotheses and to estimate the height of buildings. Mohan and Nevatia (1989a) use perceptual organization to detect and describe the buildings approximated by rectangles. The primary steps are: (1) to detect linear features and group them into parallels, (2) to form a U structure with parallel collation and aligned endpoints, and (3) to form a rectangle hypothesis with two U structures. Subsequently, they extended the results by using continuity and proximity properties to detect co-curvilinearity (Mohan and Nevatia, 1989b). Venkateswar and Chellappa (1990) proposed a framework for interpreting buildings from aerial images. First, they extracted line segments from the image and used orthogonal and proximity properties to detect potential vertices, which were then grouped together to generate edges. Then the edges were grouped to form edge-rings according to the assumption that each edge is indeed a roof edge and is searching in multiple contexts for roofs. Finally, shadow analysis of closed-rings was used to identify roofs and heuristics were used to predict that possible closes of free-rings may indicate a
new roof-edge. Lin et al. (1994) assumed the shapes of buildings to be a single or a composition of rectangular parallelepipeds and they used shadows information to form, select, and verify hypotheses from monocular views of arbitrary aerial scenes. They applied line segments to get boundaries of linear structures and computed line structures as well as relevant junctions among them. Then, because parallel relationships and parallelograms were formed for building hypotheses, promising parallelograms were selected and verified as building structures. Finally, if available, shadow information was used to help form, select, and verify building hypotheses. Moreover, shadow information with a derivation of sun angle from a camera model could be used to estimate the height of the buildings and to model 3-D building structures. They also used shadows and walls to help form and verify the hypotheses generated by the grouping process (Lin et al., 1995).

At Carnegie-Mellon University, Irvin and McKeown (1989) described a set of techniques to automatically extract and analyze shadow information in monocular aerial imagery. Four detection and evaluation systems were used: BABE (Built-up Area Building Extraction), SHADE (SHAdow Detection), SHAVE (SHAdow Verification), and GROUPER. BABE is based on a line-corner analysis to generate corners and to form building hypotheses; SHADE is based on a shadow analysis to produce the shadow region; SHAVE analyzes the building hypotheses by giving the sun illumination angle to delineate the shadow; and GROUPER is designed to cluster fragmented building hypotheses by examining their relationships to possible building/shadow edges.
McGlone and Shufelt (1994) include projective geometry in BABE to constrain a set of possible building hypotheses.

Using an expert system, Matsuyama and Hwang (1990) built a knowledge-based aerial image understanding system, called SIGMA. The objective of SIGMA is to provide a structural description of an aerial scene and to analyze the scene a single aerial image at a time. The system has four basic modules: Model Selection Expert (MSE), Low Level Vision Expert (LLVE), Geometric Reasoning Expert (GRE), and Question and Answer Module (QAM). The system works in iterative cycles. The first, or zero cycle, is a bottom-up process. MSE activates LLVE to find all possible instances of appearances from the world model. Once this basic data exists in the Iconic/Symbolic database, GRE starts the top-down cycles. This is done by building hypotheses which are based on the world knowledge of objects and their relations, and by activating MSE in order to verify these hypotheses. The final interpretation is the network which exists in the Iconic/Symbolic database after all hypotheses have either been verified or rejected, and no more hypotheses can be generated by GRE. QAM provides an option for the user to specify the particular goals and to analyze the results from the system operations.

In a more recent study by Haala (1994), the range and image data are used to detect buildings in aerial images. The steps used are as follows. First, the range data, gained from either an airborne laser scanner or feature-based stereo matching, are segmented and are mainly used to detect candidate buildings. Then, the perceptual grouping is adopted for the verification and localization of candidate buildings by using collinearity, parallelism, and symmetry properties.
Recently Gruber et al. (1995) combined a two-dimensional GIS with aerial images to create a detailed three-dimensional model of buildings. They begin by detecting line segments in the digital image. Then they projected the map detail into the digital image by using the exterior orientation of the camera. They made use of affine matching to compare two token sets based on the estimation of building height. Finally they focused on the correspondence of the specific details of buildings between the building model and the photo-texture by using either shape-from-X (e.g. shape-from-shading, shape-from-texture) or symbolic computation.

2.2 Building Recognition From A Pair or Multiple Aerial Images

At CMU, Shufelt and McKeown (1993) extended the four detection and evaluation systems from Irvin and Mckeown (1989) to extract buildings in both a monocular and stereo aerial image data set. Roux and McKeown (1994) used object space geometric constraints to reduce a large search space of possible matches for detecting and extracting buildings. They extracted building corners from multiple views of the scene by using the BABE method and then used epipolar, height, and orientation constraints for matching the corners between multiple views. The 3-D corners are linked with strong and weak image intensity gradient criteria to generate 3-D segments. Finally, parallelogram and radiometric considerations are applied to the 3-D segments to select the possible buildings.

Haala and Hahn (1995) also presented a fusion technique to detect and reconstruct buildings. First, Digital Height Model (DHM) is used to focus on the location of the
buildings. Second, 3-D line segments are extracted from the pair images by using stereo matching of gray-value edges. Third, proximity, parallelity, and rectangularity are applied to group the linear segments. Finally, primitives are matched by using least squares adjustment to reconstruct the buildings if they have such similar attributes as length, width, position, and orientation between building models and observed linear segments. Another fusion technique was described by Kim and Muller (1995). First a Canny-Petrou-Kitter filter is used to detect edge elements. Then, a line relation graph is applied to generate building boundaries by finding closed loops in the graph. Third, stereo pyramidal matching is used to get height information. Finally, the kriging method is used to interpolate the height of the building boundaries from the height information.

Dang et al. (1994) employed perceptual organization and surface models to detect and reconstruct buildings in stereo aerial images. First, they detected contours by using Canny-Deriche edge detector and estimated disparity by area-based matching. Second, planar models of surface are used to correct the sparse and noise disparity. Finally, from polygonal edge chains, linear structures, right angles, and rectangular structures are sought just as buildings are detected.

Using multiple images and a site model database, Collins et al. (1995) developed a system to acquire, extend, and refine 3-D building models and to provide a digital elevation map of the surrounding terrain. The building model acquisition involves several subtasks: (1) straight line extraction, (2) building detection, (3) multi-image epipolar matching, (4) constrained, multi-image triangulation, and (5) projective intensity mapping. In addition, with model-to-image registration techniques, the building model
extension is used to find unmodeled buildings (unexplained findings) in new images, which are then added to the site model database. In addition, a correlation-based stereo algorithm is used to extract terrain information, to provide digital elevation map, and to identify building occlusion boundaries.

Based on correlation techniques, Berthod et al. (1995) used a multiple-baseline/multiple-algorithm method to generate Digital Elevation Model (DEM) from image sequences. DEM is segmented into terrain-DEM and building-DEM by considering local height differences among neighboring surfaces. Moreover, it is assumed that buildings are represented by simple geometric shapes (squares rectangles) and are constrained by the limits of the polygon. Therefore, building-DEM is refined by polygons.

In another recent study at Bonn University, Weidner and Forstner (1995) make use of high resolution digital elevation models (DEMs) to extract buildings from digital imagery. Three steps are introduced to approach the building extraction. First, the high resolution of DEMs is generated by using the MATCH-T software package. Second, buildings are detected by thresholding the height difference within DEM. Third, the reconstruction of the buildings depends on the complexity of the detected buildings, which are divided into two types, parametric models for simple buildings and prismatic models for complex buildings or blocks of connected buildings. Forstner (1995) presents geometric models for building extraction with several elementary constraints, such as incidence, collinearity, and symmetry. To recover the 3-D building shape information, the author assumed that the geometry of an object can be described by a set of parameters.
and geometric constraints. Non-contradicting hypotheses and an independent subset of hypotheses are checked and can satisfy with the global consistency.

2.3 Conclusion on Building Recognition

Building recognition from aerial images is a complex machine vision problem since the characteristics of buildings are difficult to model and buildings present a great deal of variability in appearance. From the above literature review, it can be seen that current systems use a number of divergent knowledge representation schemes, which include constraints (algebraic, geometric, symbolic), graphs, and rules. These schemes are based on perceptual grouping, shadow analysis, corner analysis, model-based matching, and a cooperative-methods. Moreover, the representation schemes require different inputs (e.g. elevation, intensity, edges, and orientation) and produce different outputs. However, it appears that no single system has won general acceptance. It is clear that a robust solution to this problem may involve both the gathering of as much generic information about building as possible and information concerning the relationship between buildings.

Currently, it appears that there is no "best" way to do building recognition, but we can program computers to try various approaches. In addition, building recognition is a complex problem because an explicit theory of the visual process has not been developed, and because the number of possible solution paths can be enormous. Nevertheless, combined domain-dependent knowledge and empirical rules will be used to develop a building recognition solution because knowledge and rules may help to constrain the
search space by using information about the nature and structure of the problem domain. While this sounds simple, it is difficult to implement. In Chapter Five, the reasoning and process strategies for building recognition in aerial images using abduction will be presented in detail.
CHAPTER 3

THE PROBLEM DOMAIN OF BUILDING RECOGNITION
– SEEING AND THINKING

The world that we view is labeled with the meaning to us. The ability of the human eye to identify objects within an image is analogous to machine recognition of the visual elements that stand out in the landscape by their size, height, color, or any other aspect that contrasts with the surroundings. Thus the human eye can distinguish characteristic elements in a background ranging from a door on a wall to a building in a streetscape. This ability to identify parts of the environment allows us to recognize the familiar as well as to appreciate the new.

Human visual perception and machine vision are two realizations of the process of seeing, one embedded in the human brain and the other in computers. Some key questions involved in the phenomenon of sight are how people perceive their everyday physical world, how they make judgments about the occupants of the setting, and how they recognize the significant factors affecting comprehension. Although the engineering designers are not trying to model biological vision, the above questions may engage CV, AI, and DP researchers who build machine vision systems.
3.1 Visual Perception

A key question in understanding visual perception is how retinal images are identified as familiar objects (such as buildings, cars, faces). The answer may involve the mind's capacity for organizing and defining the world. In human visual perception, various cognitive processes (e.g. perceiving, learning, forming attitudes) operate to produce in an individual a vast repository of temporal knowledge of their environment. Here, perception refers to the processes that occur in the mind. These processes convert sensations into a representation of the world that we can make sense of. Sensations refer to the raw physical data which are imposed upon our senses. Although our senses often perceive partial and sometimes erroneous information, they can be supplemented and corrected by the gathering of further information about the sides of objects and about the future behavior of objects that we may not consciously notice (Neisser, 1987).

Nevertheless, it seems that human sensory systems simply reflect the physical world and our senses give us the most useful information without overloading us with information about all the stimuli. In fact, the most important property of visual perception is that under normal circumstances the perception of the local environment is immediate, effortless, and trustworthy; otherwise, we might not be able to survive on the earth.

Visual perception is a very complex process that involves not only the analysis of visual stimuli from the outside world but also an interaction between these stimuli and the viewer's knowledge about the world. Not only was human vision discussed extensively in the works of the ancient Greek philosophers but it continues to be intensively investigated by modern philosophers, psychologists, physiologists, psychophysicists,
neuropsychologists, computer scientists, and engineers. Naturally, researchers who are from different disciplines and who have made significant contributions to these subfields view the vision problem from different perspectives. A very accessible survey of contemporary theories of visual perception can be found in Gordon (1997).

As in other fields of science, the growth of empirical knowledge has led to the development of new theories and models of vision. Some examples of these new theories follow. According to the Gestalt theory, humans impose organization on their sensory perceptions. This perceptual organization is a top-down process in which grouping is guided by the characteristics of the whole perceived object (Lowe, 1985). Second, some neurophysiological evidence (such as human perception of Kanizsa figures do not require focal attention and the figures are derived from parallel process at early stages of the visual system) is supported by the findings of Davis and Driver (1994), who use a visual search task to distinguish between early and late stages in the processing of visual information. A third recent development was Gibson (1966) who presented a psychophysical theory of perception largely based on texture gradients and visual flow. For example, in the context of vision he suggested that there are many invariances in the pattern of light projected to the eye (the optic array). An invariant is some aspect of the optical array that remains unchanged with the movements of the observer in the environment. According to Harris and Humphreys (1995), both psychophysical research and neuropsychological research support the fact that the lower levels of the human visual system contain many separate channels and that the output of these channels is integrated by the higher levels of the system. Marr (1982) proposes a computational
theory of vision. Vision is analyzed as a series of representations, starting from some camera-derived image of a scene and ending with a description of the objects that are created by the image. In Marr's vision paradigm, a great deal of bottom-up numerical processing must be performed before objects can be recognized. Furthermore, he suggested that it is not necessary for machine vision to copy the actual human actual neuronal mechanism. However, machine vision can be represented by some mathematical processes that could perform functions equivalent to the biological mechanisms.

Investigations of every aspect of the visual system are necessary to lay a foundation for the study of the highest mental functions. If the signals at the input stage of the system are not understood, it follows that the final stages of the system cannot be accurately investigated. In this research we focus on developing a machine vision for object recognition. Two issues on vision are discussed: low-level processing and high-level processing, as shown in Figure 3.1.

![Generic Visual Processing Model](image)

**Figure 3.1.** A Generic Visual Processing Model (Adelson and Bergen, 1991)
3.1.1 Issues on the Theory of Low-Level Vision

The main process of low-level vision is the translation of retinal images into neural representations of the visual world. It is assumed that the low-level visual process makes available a useful, segmented representation of the two and/or three-dimensional structure of an image. Furthermore, even if our perceptual world was a world of appearance, rather than reality, it would not be a world of absolute appearance (e.g. depending on viewing position). It would be a perceived world whose characteristics are systematically related to the real characteristics of the real world.

A certain capability to capture some basic structures of reality, mostly relevant to survival, must be born by nature. For example, the human capability to distinguish shapes and colors, to recognize similarities, or to find some experiences more useful than others, could not be gained from learning alone. Such capabilities are presupposed by any process of human learning and are defined as low-level vision here. Moreover, vision researchers recognize that Marr's vision paradigm is one of the most interesting proposals concerning the earliest visual processes. The paradigm is described as follows.

Marr (1982) assumes that vision is an information processing system in which the representation of external objects is built up sequentially because the brain cannot directly infer real-world objects from the excitation of photoreceptors. Marr suggests that there are many forms of knowledge represented at varying levels of detail and that these levels begin with the reception of raw images. As shown in Figure 3.2, he uses such terms as primal sketch, 2.5D sketch, and 3D model to describe the computational process. While
these stages provide a theoretical explanation of combining primitive visual features for surface recovery, none of them specifically indicates how visual cues are combined.

![Scene Description](image)

3.1.2 High-Level Visual Processes

High-level visual processes complete the job of delivering a coherent interpretation of an image. They determine what objects are present in the scene and interpret the interrelations of those objects. Here the word high-level vision implies the general activity of the mind which may be engaged in thinking about the external of something already learnt (or memorized) and describing (or indexing) it from memory. Because visual information may come from different levels of abstraction with different degrees of precision, with or without errors (including noise), and with different degrees

Figure 3.2. Marr's Vision Paradigm (Marr, 1982)
of relation to the knowledge acquisition stage, the abilities to learn, to adapt, to modify, and to apply are truly remarkable, and are no doubt central to human intelligence. The abilities allow us to survive the changing circumstances and to develop a great variety of skills in an almost unlimited number of specific domains.

The role of top-down processing in human vision is hotly debated. However, Pinker (1985) states that there are three key questions at issue in top-down processing: (1) What kind of knowledge is brought to bear on recognition?; (2) How can that knowledge rapidly retrieve among an enormous selection of visual categories in memory (especially in long-term memory)?; and (3) How does the knowledge handle expected and unexpected objects? For object recognition with machines, some considerations are proposed, such as how to represent objects and how to recognize objects in database (or memory). Detail on reviewing object recognition is presented in Section 3-2. Here only one example, Ullman's visual routines, is introduced.

Ullman (1984) suggests that our visual systems may execute a universal set of "routines" composed of simple processes operating on a 2.5-D sketch, e.g. tracing along a boundary or filling in a region. Once universal routines are executed, their outputs could be characterized by the prominent entities in the scene, e.g. their rough shape and spatial relationships. Then this characterization could be used to trigger the execution of routines specific to the recognition of particular objects or classes of objects. The function of high-level visual processes is to determine what objects are present in the scene and to interpret their interrelations, as shown in Figure 3.3.
Higher Level Components

\[ \uparrow \downarrow \]

Routine Processor (boundary tracing, indexing, sequences of elementary operations such as the shift of the processing focus)

\[ \uparrow \downarrow \]

Incremental Representations (shape properties and spatial relations among objects and parts of the objects)

\[ \uparrow \]

Base Representations (such as primal sketch and the 2.5D sketch)

Figure 3.3. Ullman's Visual Routines (Ullman, 1984)

3.1.3 The Combination of Bottom-Up and Top-Down Processing

The dominant computational paradigms of machine vision involve the transformation of raw sensory data into some meaningful scene descriptions with logical steps that progressively make more and more abstract representations of the original scene (Fischler and Firschein, 1987). In the 70's, researchers focused on top-down processes (e.g. recognizing line-drawing objects); in the 80's, researchers shifted their emphasis to bottom-up processing (e.g. edge detection and grouping). More recently, the integration of top-down and bottom-up processing has become a major concern. For example, Desimone et al. (1995) suggest that objects must compete for attention and processing space in the visual system, and that this competition is influenced by both automatic and cognitive factors. The automatic factors are usually described as pre-attentive (or bottom-up) processes and the cognitive factors as attentive (or top-down) processes.
The bottom-up (passive) approach focuses on discovering what kind of information about the world can be initially extracted from the image. That is, this approach looks at the sensations we obtain from the environment and examines how we use the sensations to create our perceptions. It is assumed that bottom-up processes deliver relatively reliable information, allowing us to concentrate on the problems specific to top-down processes. However, the computational cost of the bottom-up processes is enormous because the bottom-up processes must be prepared to index and match virtually every object that is represented in the entire database (memory). The top-down (active) approach emphasizes the expectations about what is present in the image. That is, this approach looks at various higher mental processes that affect the way we perceive things. Although this approach also needs some searching, it can narrow down to the objects most likely to be seen with less computation. Even if we assume that the top-down processes deal with reliable representations, a job of interpretation is formidable.

It is generally accepted that both approaches require a system with a certain amount of knowledge about how the world is structured and what kinds of objects are commonly found by the distinction between various properties. Nevertheless, currently no single visual perception theory has gained universal acceptance; perhaps the main challenge is that many different factors can change our perceptual processing. And dealing with those different factors requires flexibility in thinking about how we come to know the world through seeing.
3.2 Object Recognition by Using Machine Vision

Object recognition is defined as the cognitive problems of perceiving and identifying objects. Visual recognition can be described as the matching of the retinal image of an object to a description or representation of the object stored in memory (Perrett and Oram, 1993). While machine vision development is still in its primitive stages, two types of object recognition, model-based object recognition and nonmodel-based object recognition, have been implemented. The present study is restricted to model-based object recognition, in which knowledge on the appearance of an object is provided by explicit models of its features (e.g. shape and size). For nonmodel-based object recognition, i.e., context-based and function-based object recognition, a brief description can be found in Schenk (1992).

3.2.1 Issues in Object Recognition

With object recognition, given some knowledge of how certain objects appear, plus an image of a scene possibly containing those objects, a report on which objects are present in the scene and where they are will be made. For machine vision, it is exceedingly difficult to reason back from an image to the original scene because the observation of two-dimensional image data in general contains inherently limited information about the properties of three dimensions. This lack of information implies that the solution may not be unique or that it does not exist.

Why is the problem of object recognition by using machine vision so problematic? In object recognition two important issues need to be answered. First, what
characterizes the features for representing objects in an image and in the real world? Second, how is the correspondence between image features and model features determined? Since features can be related to global or local properties, recognition is accomplished by searching for a correspondence between certain features of the image and comparable features of the model. The assertions made in the following two sections, which discuss representation and recognition in object recognition, are mainly based on research and reviews by Pope (1994), Besl and Jain (1985), Lowe (1985), Perrott and Hamey (1991), Sueten et al. (1992), Mundy (1995), Hebert et al. (1995), and Ponce et al. (1996).

3.2.2 Representation

Before the information obtained from a visual scene can be searched for recognition, it has to be represented in some form. Two representation schemes are generally used for model-based object recognition: one to represent the model of an object, and the other the content of an image. To facilitate the searching of a match between the model and the image, the two representations should be closely related. Furthermore, it is accepted that different visual tasks may require different types of representations. What makes a good object representation? Pope (1994) states that several important considerations should be taken before a model-based recognition system is built up.

1. What qualities do primitives need to represent objects?
Most approaches of model-based object recognition deal with objects only in terms of their shapes without additional detailing properties such as color and texture (Sueten et al., 1992). Similarly, images usually are described in terms of the visible manifestations of the object shape. Saund (1990) points out that the shape primitives should make explicit whatever information which is required for the task at hand and that the shape primitives should reflect the regularities and structures of the external world. Both of the criteria lead to the conclusion that shape primitives ought to depend closely on the nature of the application: what task is being performed and in what environment is the task being performed?

In this research, shape information is a major source for identifying objects, and an object is partitioned into discrete shape primitives according to three qualities: location, scale, and the category of an object at a particular location and scale.

2. The choice of coordinate system

Before the relative location of various shape primitives are described, a shape representation must deal with some coordinate system. There are principally two ways to define this coordinate system for the representation of 3-D object: object-centered and view-centered representation.

(1) Object-centered coordinate system

An object-centered coordinate system is one in which the shape primitives of an object are specified in relation to the main axis of the object itself. Usually, an object-centered representation yields the most concise and accurate object descriptions. Additionally, such object-centered description is economical because it is
valid for all possible vantage points of the observer viewing (Schenk and Toth, 1992).

However, when an object-centered representation is used, it is impossible for us to face the difficult problem of reconstructing a 3-D structure from 2-D images such as self-occlusion and perspective from the projection of a 3-D object onto 2-D image.

(2) Viewer-centered coordinate system

A viewer-centered coordinate system is one in which the shape primitives of an object are described by using the frame of a reference or coordinate system based on the observer viewing. Obviously, such representation depends on the vantage point of the observer. If more views are used, each view can cover a smaller range of objects and more accurate description. Therefore, such a system seems uneconomical because it requires a large number of view-specific aspects of objects to be stored in memory before the objects are recognized. Nevertheless, there is some interesting evidence that the human visual system uses a viewer-centered representation for object recognition (Ullman, 1989). For example, humans can recognize objects accurately and rapidly when the objects are seen from particular viewpoints and this may imply that those views of an object are readily available while others must be computed as needed.

More recent research indicates that the combination of object- and viewer-centered representations is more useful. For example, the viewer-centered description is first used to search a quick, approximate match and this match is then verified by using the object-centered description. Since the views suggest possible matches only and each
view only has to contain a few significant primitives, this approach require relatively few characteristic views (Hebert et al., 1995).

3. Types of shape primitives

Three general types of shape primitives are commonly used in the task of object recognition:

(1) Patches

Patches are chunks of data which are chosen in a simple way and require little computational cost. For example, a 2-D curve, such as an intensity edge, can be approximated by a series of straight line segments, or can be partitioned into curve segments. A 2-D surface can be approximated by a series of polygonal faces. To allow stable descriptions of a larger class of shapes, the set of primitives can be extended by adding higher-order analytic curves or surfaces.

(2) Parts

Although decomposing an object into parts is often used when parts are invariant and easily detectable, there is no general agreement on the concept of parts. However, with a limited set of parts, one can construct a large variety of objects, especially if each part can be customized by a set of parameters and the parts can be joined in several ways (Hoffman, 1995; Biederman, 1985). For example, parts can be simply chunks of 2-D lines, curves, or 3-D surfaces that can be found reliably in the image, in which the parts are visible irrespective of sensor viewing direction. In addition, generalized cylinders, generalized cones, and superquadrics are three parameterized families of parts used to describe volumes.
(3) Significant features

A significant feature is often chosen for a particular application because occasionally it will have a strong confirming effect which can positively identify a given location. In most cases, significant features will not be present. Normal procedure is to gather a set of features until some satisfaction level is reached by accepting hypotheses. However, it is important to notice what size or scale will be presented for the primitives. Regardless of the types of primitives, the significant features must be available in a sense (a range of scale) if they are to make explicit all important features of an object.

4. Organizing primitives

There are two organizing principles of relations used in computer vision but the two principles are not entirely separate. The first principle includes the ideas on part/whole relation, which is merely based on a single primitive relation of part to whole. The second principle deals not only with vertical relations between parts and their wholes, but also with the adjacency (lateral) relations among the parts of a single whole (Ballard and Brown, 1982; Winston, 1992). Both part/whole and adjacency organizations are often represented as a graph in which nodes denote primitives or groups of primitives and arcs denote the relations among them.

5. Viewpoint invariances

When an object is identified under varying conditions of poses and lighting, it is helpful to have primitives that are at least somewhat invariant with respect to conditional changes. And, the most common methods of object recognition are the
invariances for the Euclidean, namely affine and projective transformations that are often used to model the image process. For example the partial-ratio for affine mappings and the cross-ratio for projective mappings (Susse et al., 1995). Moreover, recent research is directed at finding invariances with the associated grouping method. (Weiss, 1993; Gool et al., 1995).

6. Parameterized models

When modeling an object for recognition, we may wish to describe an object whose shape can vary within certain limits or a class of objects whose shapes may differ in certain ways. One approach is to use a parameterized model in which free variables are used to specify certain measurements. The model may include constraints limiting each parameter to a set of reason values and specifying relations among parameters (Hebert et al., 1995). As more and more parameters are used, the model describes complicated objects with fewer and fewer primitives. For example, instead of using hundreds of triangles to faithfully capture a shape, a model can be approximated to describe this shape by using algebraic surfaces.

3.2.3 Recognition

The task of recognizing an object is to search for a match between a model and an image. The inputs of this task can be a library of object models. The outputs specify the identity, pose, and certainty of any objects recognized in the image. The task of recognition is made difficult by several factors: (1) not knowing what pose an object might assume in the image. (2) multiple objects and extraneous features present in the
A good strategy of recognizing an object is first to seek the features that are strong, most robust, and least costly. To recognize a single object in an image, the following sequence of steps are taken: feature detection and segmentation, perceptual organization, indexing, matching, and verification (Lowe, 1985; Pope, 1994).

1. Feature detection and segmentation

The intensity discontinuities stem from different physical phenomena: the differences in irregularities of the surface structure of objects, photon noise (quantum effects), surface orientation, and depth. Human beings can easily find contours, differentiate textures, detect and track objects, perceive depth, and analyze image contents. According to Marr’s theory on vision, the detection of sharp changes in the image luminosity is an ill-posed problem of early vision, which does not require a knowledge based approach. Moreover, the degree of the ill-posed problem will depend on the amount of the noise present in digital images (Marr, 1982; Landy and Movshon, 1991).

For the past three decades, many image segmentation techniques have been proposed (Pal and Pal, 1993). The segmentation techniques can be categorized into five classes: clustering, boundary (edge) detection, region growing, texture, and motion (Ballard and Brown, 1982). However, there are no general algorithms which will be suitable for all images. To build a general object recognition system would require the
representation and storage of a vast amount of knowledge; that is, one must go beyond simple heuristics and introduce a priori knowledge about the image contents during segmentation.

2. Perceptual organization

Human visual perception and perceptual organization refer to the abilities to organize and interpret visual sensory information. The neurophysiological approach to vision reveals us that the image can be decomposed into simple local features, such as edges, corners, and some depth information. Such low level descriptions must be organized into larger perceptual structures. For example, meaningful curve descriptions are derived from the list of edge points. To detect groupings and structures in images is believed to be the input for object recognition and image understanding.

The Gestalt psychologists highlighted a number of factors which they considered to be important in the perception of structure. These factors are (1) continuity, the preference of the human visual system for spatial continuity rather than abrupt changes; (2) proximity, a measure of the 2-D spatial separation among individual dot points in a grouping; (3) similarity, the assumption that the objects on the same surface are likely to share some characteristics which other objects do not; (4) closure, the preference of the human visual system for closed rather than open shapes; (5) symmetry perception, the existence of natural and man-made symmetry in the real world; and (6) figure-ground separation, the capacity of the human visual system for imposing a foreground-background depth organization on neighboring groupings (Lowe, 1985; McCafferty, 1990).
These grouping factors became known as the Gestalt Laws of Organization, which supposedly described why groupings occurred. The psychological explanations given for the phenomena of the grouping factors have been primarily descriptive rather than functional; that is, Gestaltists do not give an adequate theory for the role which perceptual organization plays in the overall functioning of the visual system (Sarkar and Boyer, 1993). Furthermore, because these groupings do not lead immediately to a single physical interpretation, this may be the major reason why perceptual organization has not been a focus of computer vision research.

3. Indexing to models

A number of different indexing methods have been proposed in the literature (e.g. graph matching, hash indexing, structural indexing). Recognition through indexing is the process that attempts to rapidly extract from a large list of object models those few object models that fit a group of image features, thus avoiding the need to try each object model in turn. Moreover, because choosing invariant representations is a natural method, indexing with invariances yields an attractively simple architecture. Therefore the indexing problem is related to the problem of well-known point-location or graph matching.

A general approach to graph matching consists of representing both the image features and the object features with graphs because graphs are a convenient way of representing features and the relationships between these features. However, graph matching is a difficult problem in itself because of its computational cost, which is generally exponential in the number of images or model features. One way to avoid
considerable computation is to introduce constraints such as the branch-and-bound constraint, which can be used to reduce the search space. Moreover, because it is rarely the case that the graphs of image features have the same size as the graphs of object features, searching the largest isomorphic subgraphs between image graphs and object graphs involves some form of combinative optimization (Ben-Arie and Meiri, 1987; Tresp and Gindi, 1990).

A different approach to indexing has been developed by using networks of neuron-like units that compute so-called hyper-basis functions (Poggio and Edelman, 1990). A model object is represented by a network in which individual units represent distinct characteristic views of the object. The input to the network is a vector of selected image feature measurements and the output represents either a normalized view of the object or a graded yes/no recognition response.

Another proposed indexing approach is used to organize the model library hierarchically. For example, each object can be decomposed into a list of subparts and because many objects share a common set of subparts, to distinguish one object from another the subpart relationships can be used. That is, the overall list of subparts grows sublinearly with the number of objects in the model library (Ettinger, 1988; Sengupta and Boyer, 1993; Zerroug and Medioni, 1995). The strategy of organizing the match hierarchy is that to recognize a subpart, coarse features are first used and then fine ones. By decomposing the search in these ways, a large search is replaced by numerous smaller searches, altogether requiring less computation.

4. Matching features
Given an image and an object model, both represented in terms of their features, we want to find a partial match between the two and estimate how the modeled object is positioned in the image. Pope (1994) classifies various matching methods according to whether they search for a solution in correspondence space, transformation space, or both. Ullman (1996) gives a similar classification called parts decomposition, invariant properties, and alignment approach.

(1) Correspondence space search

Correspondence space is the space of matches, which are sets of pairings between the model and image features. Objects are decomposed into parts and then parts are structured in various ways (e.g. symbolic structural description or feature hierarchical structure) to be compared with template structures. Graph or tree matching techniques are often used and the solutions generally involve a table that is assigned with information on attributes either by individual features (e.g. shape or surface) or by small groups of them. During searching, features chosen from the image are used to index the table, thus producing hypotheses about what features are present in the image. Each hypothesis denotes a possible, partial match between the model and image, which must be further tested to determine the full extent and quality of the match. How well matching is done may depend on how well entries are distributed throughout the table. Matching performance is also greatly affected by uncertainty in feature measurements.

(2) Transformation space search

Transformation space is the space of possible object poses, viewpoints, or transformations between an object and a camera. It is assumed that certain
properties of objects are invariant and still can be traced under transformations (e.g. scale, translation, rotation). For example, the Hough transformation is a common and efficient technique for detecting lines and curves in a 2-D space. The other transformations such as affine, projective and log-polar transformations are also often used in transformation space search.

(3) Using both search space

This includes a two-step process. The search begins with correspondence space where it matches just enough anchor features (or most possible pairings of features) to determine viewpoint transformation. Once the viewpoint transformation is determined, it is used to project the remaining features of the model into an image for the verification of matching. This approach uses a set of views on an object instead of its model. Furthermore, it also requires at least one set of anchor features to produce a sufficiently accurate estimate of the viewpoint transformation (Ullman and Basri, 1991; Pope and Lowe, 1996).

5. Verifying the match result

There is relatively little research on how to make a decision on whether an optimal match has been found. Many systems simply require that the fraction of the features of a model be matched, and/or that some fraction of the edge projected from the model lie near image edges (Chen and Kak, 1989). Match solutions are verified by testing these measures against empirically determined thresholds and are then ranked according to the measures to select the best, mutually-consistent solutions.
3.2.4 A Note on Object Recognition

Human object recognition (or identification) is remarkable not only for its flexibility but also for its speed. For example, humans can identify familiar objects in about 800 milliseconds (Dodwell, 1995; Bruce and Green, 1985). Furthermore, the response time for a given scene is the linear dependence on the amount of objects received on the time necessary for recognition (Bullock et al., 1994). For machine vision systems, to recognize any number of objects in an image, most methods are used to recognize a single object repeatedly until no further objects are found or until there is nothing more to explain. Although more recent systems can handle larger model bases, we still do not know how to formulate the constraint and the context that a model base places on a grouping, how to formulate the usefulness of a given activity in the context given by the model base, or what is known about the scene (Mundy, 1995; Hebert et al., 1995).

Currently there seems to be no universal object representation for all aspects of application as well as no practical, robust searching method which can index and match the objects in a 3-D nature scene based on the observations obtained from few intensity views. However, Ponce et al. (1996) suggest that both the strongest model applicable to the task and the multiple cues or representational levels must be used in order to facilitate object recognition tasks. A good review of model-based object recognition techniques may be found in Pope (1994).
CHAPTER 4

KNOWLEDGE REPRESENTATION AND ABDUCTION

Reasoning with visually perceived information is a widely known aspect of intelligent behavior. The gap between the human ability to process visual information and the current level of machine vision is enormous. For example, although aerial images include mixed-feature environments, people still can identify features from the images but machines can not. This inability is not because there is something wrong with machine vision technology, but because we know too little about how the human eye/brain works.

4.1 Knowledge

"Knowledge" can be difficult to define, but primarily it is a cognitive notion: what we "know" is anything that is consciously at our disposal. Knowledge is important because almost everything that we do is in some way based on the knowledge which we have stored in our brains. The ability to use knowledge consciously is surely one of the most fundamental skills that set humans apart from the other members of the animal
kingdom. However, people can only use the knowledge that they have, i.e., they are not able to execute tasks with which they are not familiar. Nevertheless, in our daily life with visual perception, humans excel at object recognition (e.g. recognizing a tree or a building). This ability appears to be based on an individual’s empirical knowledge (e.g. common sense knowledge) and/or rational knowledge (e.g. domain dependent knowledge) to recognize objects in the surrounding environment.

Although psychology, philosophy, and related disciplines offer various perspectives and methodologies for studying intelligence, defining intelligence is also problematic because intelligence appears to be the capability to mix multi-information processing and representation (Pylyshyn, 1984). Moreover, the theories proposed in such fields as psychology and philosophy are too incomplete and too vaguely stated to be realized in computational terms.

However, it is believed that knowledge representation should not only play a central role in computational modeling but is also very important to building recognition. To further the development of machine vision, the major questions to answer is to what extent can humans recognize buildings from a very general template or a small number of them, and to what extent having previously seen a variety of buildings allows us to use an approximate building form to recognize new detailed forms.

4.2 Knowledge Representation

Humans recognize buildings from aerial images unconsciously, and this process is not actually understood. Knowledge sources that humans take for granted are shown to
be critically important in our understanding of what we see. The interpretation of images is not a simple process for machine vision or human vision. We are not aware of the abundance of special visual cues that we use to make sense of images. For example, obvious cues, such as color or shape, and subtle cues, such as shadow or texture are important in distinguishing objects. Each of these is a source of evidence about a scene.

But how can different kinds of cues lead us so quickly to identify objects in complex environments? How can one make changes in case of error or if new evidence is found? How does one resolve inconsistencies? How does the vision process exploit knowledge associated with general, nonvisual activities? Indeed, there is much about the vision process that we do not know.

How knowledge is to be represented depends on what knowledge needs to be represented and how it is going to be used. The ability to represent and use a body of knowledge or information is fundamental to all problem solvers. It is well recognized that different problems require different representations so that the corresponding algorithms can be implemented efficiently (Brachman and Levesque, 1985). Thus, a representational specification should ideally consist of the content of the representation, how it is structured, and what kinds of operations it supports. For building recognition, machine vision has many alternative representations, most of which manifest themselves as a set of constraints that are applied to parts of an image. Each constraint guides the interpretation locally, and hopefully can be combined with all other constraints to produce a global analysis of the image.
It has long been hypothesized that perception is a form of inference (Charniak and McDermott, 1985). As Marr (1982) states, "Although the real world is three-dimensional, we perceive it as such, each retina is only two-dimensional. Since the mapping from the world to the retina is many-to-one, the possible states of the world consistent with a retina image are many. The result of all this is that knowledge of the world is inferred. Inference lies at the heart of perception." Furthermore, on the basis of visual perception research, the notion of schemata or frames in guiding perceptual activity has been used in several theories to emphasize the role of expectations (Marr, 1982; Ullman; 1984; Pinker, 1985). According to such theories as hypothesizing a specific object and then using detailed prior knowledge about this object in an attempt to confirm or refute the current hypothesis, we propose the building recognition via one of the logic inferences, abduction. Although there has been substantial research in abduction, especially in the past ten years, few studies have focused on object recognition. The goal of the present study is to attempt to find the "best" set of hypotheses which will allow us to recognize buildings from aerial images.

4.3 Abduction

Abduction is a logical inference, a pattern of justification, or a type of reasoning used in explanation finding and a variety of consequence finding. In many cases, based on limited knowledge, humans are able to generate an explanation to account for some particular fact or event. This process of explanation generation may be abduction, as Charles Sanders Peirce (1839-1914) describes it: the generation of hypotheses to explain...
observations or conclusions. Furthermore, Peirce noted that abduction is a companion process to deduction and induction. Fann (1970) gave an example to show the difference among three logical inferences.

Deduction:
premise: All the beans from this bag are white.
observation: These beans are from this bag.
conclusion: These beans are white.

Induction:
premise: These beans are from this bag.
observation: These beans are white.
conclusion: All the beans from this bag are white.

Abduction:
premise: All the beans from this bag are white.
observation: These beans are white.
conclusion: These beans are from this bag.

From the above example, we can see that deduction is truth preservation; induction is an inference from a sample to a whole; and abduction is an inference from a body of data to an explaining hypothesis. Peirce suggests that abduction plays the role of generating new ideas or hypotheses; deduction serves to evaluate the hypotheses; and induction is the justification of the hypothesis with observational data. If the abduction is correct, it is a way to generate new knowledge.

Pople (1973) explored abductive reasoning in AI early on, and he was mainly concerned with using abduction to perform disease diagnosis. Since Pople, various
applications of abduction have been reported. Chamiak and McDermott (1985) described abduction as a general model for explanation and recognized that many diverse AI tasks, including natural language understanding, diagnosis, some kinds of default reasoning, and image interpretation, can be elegantly modeled as abduction.

Generation, evaluation, and deciding whether to accept (or comparing with alternatives) are all parts of abductive reasoning. The whole process is required to generate an abductive justification, that is, a justification of some conclusion as "best explanation" (Josephson and Josephson, 1994). In the ideal case, new consequences are deduced from a hypothesis and tested against additional data. Nevertheless, abductive reasoning is fallible. For example, in Fann's example, white beans may come from other sources. Therefore, hypotheses can't be held up as valid simply by looking at how they are abduced.

4.3.1 Josephson's Account of Abduction

The simplest pattern of abductive reasoning is to infer the conclusion "H is plausible" from the two premises "if H then D" and "D is observed to be true", where H stands for a hypothesis and D for a collection of evidence. Clearly, the most important characteristic of abductive inference is the generation of hypotheses, which, if true, would explain some collection of observed facts.

Currently, the description of abduction such as "inference to the best explanation" is accepted by most researchers on abduction. Then, the central problem of abduction is to
understand the conditions or the criteria for the “best hypothesis”. One of the examples of inference to the best explanation is Josephson’s account of abduction.

The schema:

D is a collection of data (facts, observations, givens).
H explains D (would, if true, explain D).
No other hypothesis can explain D as well as H does.

Therefore, H is probably true.

The criteria:

The judgment of likelihood associated with an abductive conclusion should depend on the following considerations (as it typically does in the inferences we actually make):
- how decisively H surpasses the alternatives
- how good H is by itself, independently of considering the alternatives
- judgments of the reliability of the data
- how much confidence there is that all plausible explanations have been considered;
- how thorough was the search for alternative explanations.

Beyond the judgment of its likelihood, willingness to accept an abductive conclusion should (and typically does) depend on:
- pragmatic considerations, including the costs of being wrong and the benefits of being right
- how strong the need is to come to a conclusion at all, especially considering the possibility of seeking further evidence before deciding.

Figure 4.1. Josephson’s Account of Abduction

(from Josephson and Josephson, 1994, page 14).
The essence of Josephson’s account of abduction, I think, is the consideration of alternative hypotheses in that with certain criteria, subsets of the possible combinations of hypotheses are examined.

One key problem in abduction is to generate and evaluate hypotheses and abductive conclusion. Fox (1992) and Schenk (1995) indicate that the abduction task typically consists of three subtasks: (1) the generation of possible hypotheses, (2) the selection of plausible hypotheses from the set of available possible hypotheses, and (3) the construction of composite explanations from selected plausible hypotheses. The three subtasks are discussed in the following sections based on the work of Josephson and Josephson (1994), Leake (1993), Rossotto and Dupre (1995), Fox (1992), Narayanan (1992), Abdallah (1995), Amedeo and Golledge (1975), Bayazitoglu (1995), Algarni (1992), Lin (1992), Kelly (1988), Ho (1994), and Holland et al. (1986).

4.3.2 The Generation of Possible Hypotheses

The word "hypothesis" suggests something tentative, an idea that has not been proven yet. There are two fundamental knowledge sources that will help us to generate hypotheses:

(1) Domain Knowledge

Although there may be millions of possible explanations to a phenomenon, humans seem to be able to reason efficiently when confronting a very complex scene. It is believed that we do not have to examine every hypothesis before we arriving at the possible one. One reason for this is that when identifying a specific
problem we may only make use of the limited subset of our knowledge relevant to the problem. Furthermore, the more domain knowledge provided to the system, the shorter the search effort. We should try to make explicit all the priori knowledge that is relevant.

(2) Empirical Rules

Empirical rules, a kind of hypothesis or heuristics, are sometimes adopted to explain happenings. They show that the observed phenomenon follows logically, either probably or necessarily. The basic idea is that of knowing associations that can help generate hypotheses, but which do not represent "domain model" knowledge. Moreover, because of the usual presence of imprecision, noise, and incomplete information in machine vision, not only domain knowledge but also some flexible heuristics can be derived in order to accept explanations that are only partially complete and/or consistent. Nevertheless, we should realize that although a heuristic is useful because it eliminates the need to express knowledge precisely, a heuristic is limited because it still fails to address the vast bulk of knowledge which exists in our brain.

There is no exact guideline we can follow to generate hypotheses. However, the human mind is akin to the truth in the sense that in a limited number of guesses it will light upon the correct hypothesis. Usually, the more familiar hypothesis will be tried first, and the preferred hypothesis is the one that reflects to the greatest extent possible a synthesis of ideas evoked from the observation. The more data a candidate hypothesis accounts for, the more highly regarded it is by the abduction processor.

Furthermore, the discovery of plausible hypotheses early in the problem domain provides a useful decomposition of the task and reduces the search space. Moreover.
abducing from observation data as much as possible may reduce the computational complexity. Other useful heuristic criteria will undoubtedly be forthcoming, but what form these criteria may take is still an open question.

4.3.3 The Selection of Plausible Hypotheses

To select plausible hypotheses from the possible hypotheses is an important step in abduction tasks. Depending on the nature of the hypothesis to be evaluated, the crucial aspect of an evaluation task is to recognize the value of a subject to the state hypothesis. The hypothesis evaluation tasks can be distinguished in terms of whether the subject selects the evidence from a large pool of information or whether he is presented with data and asked to determine the relevance of the data. Most of the work on hypothesis evaluation requires the subject to select the data that he wants to evaluate. However, there are tasks in which the subject is presented with certain data and asked to determine what the data indicates with respect to the truth or falsity of the hypothesis.

More generally, to evaluate a hypothesis we may ask if it is adequate. An adequate hypothesis should satisfy the following criteria:

1. Relevance: the influence of the hypothesis on observation data should have a high degree of relationship,
2. Completeness: the hypothesis should explain the observation data as thoroughly as possible,
3. Consistency: the hypothesis is not incompatible with other knowledge,
4. Testability: the hypothesis must be capable of empirical verification, and
(5) Simplicity: the hypothesis is chosen if there is more than one hypothesis equally satisfied.

Thus the following two steps can be taken to determine the relation of the hypothesis to the current case: (1) scoring the hypothesis by using some predefined quantity measurements to determine plausibility (e.g. Bayesian statistical analysis), and (2) testing the hypothesis by examining what data items the hypothesis can explain in the context of the current case and comparing the result with predefined criteria.

A hypothesis consists of a set of relevant instances for determining the true value of the hypothesis. A common strategy among researchers is to select and test the data supporting the hypothesis and to exclude the data conflicting the hypothesis. However, under certain conditions, when looking at the value of the excluding data, we can improve the common hypothesis test strategy and tend to have a more reliable hypothesis procedure. Moreover, we often ignore the possibility that the given information can be consistent with alternative hypotheses and instead tend to overemphasize the value of the fact.

4.3.4 The Combination of Plausible Hypotheses

Abductive reasoning is similar to the phenomenon in which people tend to look for evidence that supports a particular belief, but not evidence that might disprove the belief. According to past experience, a certain hypothesis is adopted because it will explain the observed facts. And selecting the hypothesis from different points of view will lead to different results.
Nevertheless, the central issue of abduction, which is determining the condition or criterion for the best explanation has yet to be understood. Fox (1992) states that the best explanation can be defined as follows.

(1) The explanation is complete or it is as detailed as possible. For example, hypotheses explain all observations;

(2) The explanation is consistent or the explanation should be internally consistent. For example, no parts of hypotheses are incompatible; and

(3) The hypotheses are minimal. For example, the least number of hypotheses contain full explanation.

However, a basic problem which naturally arises is that there may be many different possible sets available. Thus it is necessary to determine which one of the preferential ordering on the explanations is best. In general, when completeness is insisted upon, it is an NP-hard problem, and no efficient algorithms are available.

Nevertheless, there are several ways to handle such computationally intractable problems. For example, adding some constraints to the specification problem will reduce the search space. Another way is to go for some form of approximation, i.e. to give up the idea of finding the best solution and be happy with a "good" solution. Adding constraints means restricting the expression of the problem; going for approximate answer may mean giving up completeness of reasoning. In abduction, some simplification is useful and necessary to facilitate analysis. For example, incompleteness, including a high-confidence explanation, is accepted as a best explanation.
4.3.5 A Note on Induction and Abduction

In logical inferences, both induction and abduction are the opposite of deduction. Besides the background knowledge, the three forms of hypotheses, induction, abduction, and deduction, still require a given set of observations. However, the two inferences, induction and abduction, differ significantly in the way they satisfy. For example, induction and abduction are important forms of hypothetical reasoning; they are inferences from essentially incomplete information. Nevertheless, induction refers to inferring data from sampling data and abduction refers to reasoning from effects to causes or explanations.

Inductive reasoning is used to make a judgment about the likelihood of the hypothesis, which is based on the accumulated evidence. And inductive reasoning is used when we want to synthesize the information conveyed by the observation data into a hypothesis that can account for all the observation data together in a common way. Abductive reasoning is used to generate an explanation for the truth of the observation in terms of hypotheses which are typically specific to the situation and individual objects at hand. And each separate observation will get its own explanation, not necessarily related to the explanations of other observations and not intended to account for other observations. Nevertheless, this distinction between abductive and inductive hypotheses does not seem to be fully captured by contemporary models of abductive and inductive reasoning in AI. A more detailed discussion about abduction can be found in "Abductive Inference: Computation, Philosophy, Technology" by Josephson and Josephson (1994).
CHAPTER 5

REASONING AND PROCESS STRATEGIES FOR BUILDING RECOGNITION: A PROPOSED GENERIC BUILDING MODEL

The perception of a physical object such as a building, seems partly determined by culture. Objects can possess complex levels of meaning which must be learned; they are not given by the object itself. We interpret our physical environment consciously or unconsciously in a variety of ways such as interpret a symbolic drawing of buildings. The building recognition concept presented in this research rests on three major assumptions. First, abstract or physical models are essential for analyzing the complex tasks involved in building recognition. Second, a building recognition hypothesis cannot be achieved until a variety of stimuli has been unified. Moreover, the first unification must be accomplished by natural intuition (e.g. depth and motion detection, edge/region detection, and grouping) in visual processing. Third, because most visual perception problems cannot be solved in a single step, the vision task must be decomposed into subtasks. In each subtask it is necessary to hypothesize and confirm a single fact or multiple facts. Among different subtasks, any hypothesis that does not violate the
constraints imposed by other subtasks can be considered a consistent or plausible description of the objects.

5.1 A Structure of the Features on the Earth’s Surface

Building recognition is a part of scene recognition. A scale defines the relationship between the measurements of the features as shown on the aerial images and as they exist on the earth’s surface. Features on the earth’s surface are divided into two general classes: man-made features and non-man-made features. As shown in Figure 5.1, man-made features consist of structures made by people or machines, such as buildings, roads, vehicles; whole the non-man-made features include natural phenomena, such as mountains, waters, vegetation.

![Diagram of the features on the Earth's surface](image)

Figure 5.1. A Simplified Classification of the Features on the Earth’s Surface

Our knowledge continuously affects our seeing; we learn not only how to see but also what to see. Among the features that make up our physical environment, how do humans recognize buildings? It is proposed that extensive knowledge about the
characteristics of buildings is a vital component of vision and may help us to recognize the buildings in a scene, even though the buildings may present a great deal of variability in appearance.

5.2 The General Characteristics of Buildings

In the Random House Dictionary of the English Language (1987), a building is defined as “a relatively permanent enclosed construction over a plot of land, having a roof and usually windows and often more than one level, used for any of a wide variety of activities, as living, entertaining, or manufacturing”. From an architectural point of view, all buildings can be categorized by types or styles. The type is a matter of physical use and social function, while the style is comprised of visual effects such as shapes, ornament, and proportion (Krampen, 1979; Weidhaas, 1985).

Before the 19th century, buildings tended to be somewhat simple structures. They primarily provided shelter from the weather and preferably were also required to be visually pleasing. Moreover, buildings generally did not exceed five stories in height and their exterior walls were provided with openings and windows for light and ventilation. By the 19th century, buildings became more sophisticated because advances in building materials and scientific methods could be applied in building design (Gowans, 1992). Although architects make extensive use of their knowledge of existing buildings or precedents, new buildings or their components seldom form exact copies of the precedents (with the exception of certain standardized building parts). This means that the concepts of building design change over time, and this change continues because
every building represents a composite idea of all the centuries of ideas that have preceded
(Weidhaas, 1985; Merritt and Ambrose, 1990).

The physical structure of a building can be described in terms of its geometric
information (e.g. the location of walls and roofs) and attribute information (e.g. colors
and materials). The physical structure of a building includes the volume that it occupies
and its spatial relationship to other objects in the vicinity. In this research, building
recognition focuses on the exterior appearance of the buildings, which is influenced by
many factors: available construction materials, climatic conditions, technological
inclinations, social and political organization, subsistence strategies, cross-cultural
experiences, religious beliefs, and world views. No single factor can dominate the
exterior appearance of the buildings. Furthermore, different scales of building
appearance will influence the visual cues on the outline of the building.

5.3 A Proposed Generic Building Model

Models of buildings are used to focus on the relevant or interesting aspect of the
real world situation and to eliminate unimportant information. The following three steps
are usually used to construct a model of building. First, a well-defined problem is
required. While this seems to be very general, it is very important. If the problem is
stated broadly or vaguely, it is difficult to precisely construct the stated hypotheses of a
model of building. Second, hypotheses are stated and assumptions are made. The
hypotheses are statements about what one thinks the principal relationships are in the
problem situation and the assumptions are used to simplify complicated behavior or
conditions in the problem domain in order that the behavior or conditions may be handled. Third, reasons are given for how the model will be constructed. Once the assumptions and hypotheses are made, some questions still need to be answered. For example, what are the relationships and/or interactions between the objects in the real world that are going to be modeled? Are the variables measurable? Are there other substitutions?

Normally, steps 2 and 3 need to be tested and revised in real world situations in order to achieve acceptable model performance. Moreover, there are three important questions that one needs to answer before actually setting up a building recognition model: (1) What kind of model is needed: a generic, a few, or full reality building models? (2) Should the model be view-centered or object-centered? (3) Should the model be based on part/whole, spatial, or geometric relationships? It seems that there are no conclusive answers to the above three questions. For example, if we set up a large number of building models, it may be relatively easy to match the appearance of the buildings between the image and model, but the search space would be very large. On the other hand, if only a generic or a few very important features are considered, it may be impossible to capture all the characteristics of the building that distinguish buildings from other things.

It is argued that surface properties, part/whole relations, and prior knowledge (e.g. view position, site location, weather condition, and image scale) are very important factors for recognizing buildings in a scene. For example, when buildings are viewed from outside, roofs and walls appear as the major architectural features of the buildings
and that the main design considerations are the accommodation of weather conditions, economic cost, and aesthetic appearance. Almost all walls are perpendicular to the ground but roofs are not. Because we can not foresee all the viewing directions of the buildings and generating models of all buildings would be impractical, an object-centered generic building model is adopted. This model is mainly based on the structure of the building form (e.g. the size and shape of the walls and roofs) and the surface properties of the building (e.g. the orientation and location of walls and roofs). Other components (e.g. window, door, and chimney) on the building parts (e.g. walls and roofs) are subparts of the building, as shown in Figure 5.2.

![Semantic Network of Generic Building Model](image)

**Figure 5.2. An Instance of Semantic Network of Generic Building Model**

The part/whole, spatial, and geometric relationships between buildings and other features in a scene are used to determine the existence of buildings. At different vertical level, the link forms the part/whole relation while the same horizontal level has the spatial
and geometric relation. The geometric and spatial relationships involve the physical dimensions and the locations in three dimensional space. For example, the spatial relations include above, below, next-to, in, and out; the geometric relations are parallel, vertical, horizontal, larger than, smaller than, etc.

5.4 Proposed Hypotheses and Assumptions for Building Recognition

Since the process of vision is an immensely complex one, theories at many different levels of abstraction must be utilized. Although the numerous investigations of the problems of vision rarely yield complete theories, their individual contributions have resulted in the formulation of constraints for constructing a theory. For example, because of an image formation view, a 3-D scene is projected onto a 2-D image. If we want to recover a description of the scene, somehow the loss of one degree of freedom must be overcome by using constraints. Additional constraints may be derived from domain knowledge (e.g. how and where the images were taken). Moreover, because the process of vision is so complex, explanations can be put forward at many different levels from assumptions necessary to solve abstract vision problems to restrictions on the machinery that will implement the solution. If constraints at any level are missing, then the visual processing will be underconstrained for recognizing objects.

Previous researchers in building extraction stressed the importance of explicitly computing representations of such qualities as intensities, illumination, depth, surface orientation. Most of the approaches they used focused on explicit hypotheses and the explanatory knowledge was made explicit. The differences between the above
approaches lie in how the explanation is generated and how the evaluation of hypotheses is performed in order to choose the "best" one. Undoubtedly, there are many ways of implementing building extraction and many new constraints still need to be discovered.

Machine vision requires sufficient criteria (e.g. partial knowledge) for recognizing buildings from aerial images and is not necessarily the knowledge of all the details. In this research it is assumed that straight lines are highly related to man-made objects. Given a pair of aerial images and the exterior orientation, we can measure the absolute coordinates of an object. If the pair of aerial images are resampled and epipolar images are gained, we can narrow down search the correspondence between the two images.

Furthermore, the domain knowledge of building appearance is represented explicitly with small sets of hypotheses, which are used to explain a set of findings of 3-D primitive features to generate the generic building model. Since a building is a form of enclosure with walls and roofs, hypotheses of building appearance we made are mainly related to walls and roofs. Based on spatial relations between nearby building features and buildings, some hypotheses are generated and used to explain the findings. For example, as shown in Figure 5.3, the hypotheses of an instance of generic building model in a suburban area are:
(Walls
  size:
    top height >= 3 meters
    width >= 3 meters
  shape:
    flat surface
  orientation:
    perpendicular to the ground
  location:
    on the hump)
(Walls_opening
  size:
    top height < 3 meters
    width < 3 meters
  location:
    on the Walls)
(Roofs
  size:
    length >= 3 meters
    width >= 3 meters
  shape:
    flat or curved surface
  orientation:
    not perpendicular to the ground
  location:
    upper surface of the hump)
(Roofs_utilities
  size:
    length < 3 meters
    width < 3 meters
  location:
    on the Roofs)
(Roads
  size:
    length > 5 meters
    width > 3 meters
  shape:
    flat or curved surface with parallel lines of surface boundaries
  location:
    on the ground)
(Vehicles
  size:
    4.5 meter > top height > 1 meter
    length > 2 meters
    width > 1 meter
  shape:
    flat or curved surface
  location:
    on the Roads or on a parking lot)
(Vegetation
  shape: irregular surface
Vegetation_tree
  location: above the ground)

Figure 5.3. The Hypotheses of an Instance of Generic Building Model in a Suburban Area
5.5 Proposed Reasoning and Process Strategies for Building Recognition

This research continues Schenk's model of layered abductive building recognition. It emphasizes reasoning and control strategies, and searches for a set of hypotheses that will enable a computer to recognize buildings from aerial images.

5.5.1 Schenk's Model of Layered Abductive Building Recognition

Based on the hypothesis that visual perception is layered abduction, Schenk (1995) proposed a concept of layered abductive building recognition model, as shown in Figure 5.4. First, under various assumptions and constraints, the object recognition problem is decomposed into different layers. Second, the object recognition begins with preprocessed images and progresses from intermediate levels of raw and segmented surfaces towards a geometric and semantic description of a building.

```
Segmented Surface
  ↑
Refined Surface
  ↑  texture, shading, color
Raw Surface
  ↑
Hierarchical Stereo
```

Figure 5.4. Schenk's Model of Layered Abductive Building Recognition (Schenk, 1995)
5.5.2 A Proposed Reasoning Strategy

As far as building recognition is concerned, the following issues are considered to decompose the recognition problems. First, the visual cues chosen to be processed is obviously an important consideration because all further processes depend primarily on how well this initial stage is carried out. Second, whether it is helpful to break an image down into its basic elements (such as primal sketch) needs to be determined. Third, is it possible to use simple local computations to extract quantities that correlate well with the global properties of a building?

5.5.2.1 What Kinds of Cues are Used?

Generally, the particulars of a scene cannot be accurately surmised until they are seen. Our visual system is equipped with the capability to describe an object by using visual stimuli (e.g. shading, motion, color, texture, and disparity) and to guess what the object is from its description. This guess is based on the findings in image space (e.g. points, lines, planes) and the explanations in object space (e.g. roof corners, roof boundaries, roof surfaces).

In object space, physical edges are one of the most important properties of scenes because they are related to the boundaries of objects, or to the changes in surface orientation, material properties, and so on (Torre and Poggio, 1986). In image space, most edges and shapes are view-point dependent and they can result from discontinuities in surface normal, in self-occlusion (where a surface is tangent to the line of sight), in surface reflectivity, or in illumination.
Recognizing a visual environment is a thing that people do extraordinarily well and it is believed that our visual system has the capabilities to describe the shape of an object and to hypothesize what the object is from its descriptions. Although this hypothesis might not be exactly correct, using shape information simply narrows the potential matches and triggers visual computations designed to narrow hypotheses further. This hypothesis is described as an abduction inference because the conclusion is a "best" interpretation of the data. However, the truth of premises (e.g. the description of a shape) does not logically guarantee the truth of the conclusion (e.g. the identity of the object), so the inference is fallible. Moreover, because the truth of the conclusion does not logically follow from the truth of the premises, the strength of the inference must derive from some other sources. These sources, we claim, come from the domain-dependent knowledge, which generates hypotheses to form the best explanation. That is, both visual stimuli and prior knowledge (e.g. area size, site location, and association) are great useful cues for recognizing buildings from aerial images.

5.5.2.2 The Use of 3-D Surfaces

Previous research on building recognition of aerial images focused on traditional computer vision problems such as segmentation, grouping, and matching (not only matching primitive features between left and right images but also matching primitive features between the image space and the model of object space) with 2-D lines, planes, and 3-D lines (Shufelt, 1996). Nevertheless, 2-D lines, planes, and 3-D lines alone are insufficient for building extraction or recognition. For example, the projections of depth
or orientation discontinuities in a physical scene result in the intensity edges of images, which are not ideal step edges but are a combination of steps, peak, and roof profiles. Currently with most of the schemes for edge detection, it is difficult to detect and localize the composite nature of intensity edges. Furthermore, it is known well that three major limitations of using edge representations are (1) ambiguous scene interpretation (e.g. caused by occlusion, shadows, surface orientation discontinuities), (2) loss of data (e.g. caused by noise or low contrast), and (3) complexity matches between image and model primitives.

Another possible way of recognizing buildings is using 3-D surface to describe buildings, rather than using only pixel-based or feature-based descriptors. It is believed that the richer the descriptors are, the more reliable the influences will be. It is generally agreed that for recognition the primitives should be invariant to rotation, translation, and scale as could as possible. Surface representations seems ideal for scene recognition since feature surfaces are what can be actually measured by stereo matching techniques or active sensor. For example, surfaces are large and stable features, and in a scene information about surfaces can be computed from 3-D depth, color, or texture such as the 3-D depth can be derived from stereo, shape from shading, shape from texture, or active range sensors. A comparison of object recognition systems according to their indexing primitives is provided by Dickinson et al. (1992). The indexing primitives range from low complexity (e.g. 2-D points) to high complexity (e.g. 3-D volumes). Moreover, because the simple indexing primitives represent a more ambiguous interpretation of the image data, they must rely heavily on a top-down verification to disambiguate the data.
Many researchers are working on surface-based recognition. For example, the importance of Marr’s proposal lies in having a reconstructed surface representation as a significant intermediate entity in the image understanding process. And Marr’s 2.5-D sketch represents mainly surface orientation, depth, and labeled boundaries with region groupings. Besl and Jain (1986) use the sign of Gaussian and Mean curvature to classify the points on the surface. Fisher (1989) describes how surface data can help overcome recognition problems, such as using surface data to help select models, dealing with occlusion, and reducing matching complexity and coping with noise. In this research, a building is described locally in terms of its surfaces and it might be in the form of a local orientation of the surface relative to the direction of a camera. Then, the reasoning is made on a description of the 3-D surface patches and facets, which are obtained by means of a surface reconstruction step or other sources. For example, the sources of surface data may come from stereo, optical flow, laser or sonar finding, surface shading, surface contour, and various forms of structured lighting.

5.5.2.3 What Hypothesis is Used?

A hypothesis is something that is answerable to evidence or data. Consider the duck-rabbit figure made famous by Wittgenstein’s discussion in his book, Philosophical Investigations. From one vantage point the figure looks like a drawing of a duck and from another like a rabbit. Since the stimulus approximates both the figures of duck and rabbit, the gestalt switch in the appearances represents a switch in the hypotheses.
Therefore, to find sufficient evidence in order to select a set of hypotheses out of any alternative hypotheses is of crucial importance.

For building recognition, we deal with approximation, incompleteness, and heuristic strategies. A solution to building recognition is a “maybe” answer, equipped with reasons for us to believe that maybe is actually a “yes” or a “no”. It is argued that even if a scene is well-understood, recognizing the possible outcomes of the scene is still computationally intractable. Thus, restricting our attention to the significant or important outcomes is necessary. This is also true of the task of building recognition.

We believe that a strong hypothesis, if it is held, allows us to explain a lot of separate pieces of evidence. This hypothesis might join other hypotheses and become a single coherent explanation that has the additional advantage of providing us with the basis for having great insight. Moreover, because a good explanatory hypothesis does not just explain the obvious data, it might lead to a unclear observation data. Can a strong hypothesis (or some hypotheses) for the task of building recognition be found? In this research, based on the unique property of aerial images, the top view of a scene, a reasoning strategy for the proposed building recognition of a pair of aerial images is to find the roofs of a building, as shown in Figure 5.5.

Furthermore, the roofs of a building are hypothesized from 3-D flat or curved surfaces with the constraints of surface properties (e.g. area size, length, width, elevation, and orientation). The boundary shapes of 3-D flat or curved surfaces may include 3-D parallelogram, parallel lines, straight lines, and corner point (intersection of two straight
Moreover, the above 3-D information is obtained from matching the 2-D straight lines and corner points, which are assumed to be related to man-made objects.

Find building
↓
Find roofs
↓
Find 3-D flat or curved surfaces
with the constraints of surface properties
↓
Find 3-D parallelogram, parallel lines,
straight lines, and corner points
↓
Find 2-D parallelogram, parallel lines,
straight lines, and corner points
↓
Raw images

Figure 5.5. A Reasoning Strategy for the Proposed Building Recognition

5.5.3 A Proposed Process Strategy

Perception is so simple a thing that we take it for granted, but actually it is a combination of many kinds of visual information such as shape, size, color, lightness, distance, movement. In terms of evolutionary process, the human visual system of the cerebral cortex has three major separate pathways: one responsible mostly for the
perception of color and brightness, another responsible mostly for the perception of fine
details of shape, and the other responsible primarily for the perception of movement and
certain aspects of distance (Livingstone and Hubel, 1988). Each pathway performs its
functions somewhat independently of the others, although in some unknown way the
brain combines the information in the three pathways to experience each stimulus as a
unified whole. However, how do we unify our perception based on the three parallel and
partly independent visual pathways? At this point investigators do not have an answer.

5.5.3.1 Determining Levels of Process

Generating explanations in complex problems within incomplete domain theory
often requires using hypotheses at different levels of knowledge because a single level of
process seems insufficient. Instead, multiple levels of process can be used to infer an
explanation, not at one level of knowledge, but at all important levels. For example, the
explanation can be function/structure or part/whole relationships. Normally there are
different entities at different levels to be processed and to pass intermediate results
between levels (Grene, 1987). The characteristics of the level are: (1) the variety of
behavior is obtained by system structure, (2) every level within a system is explained
only in accordance with the mechanistic laws specific to its level, (3) each level must
have an internal coherence, and (4) one level is higher than another if the entities at the
higher level are composed of the entities at lower levels (Bainbridge, 1993; Griffiths,
1995).
Because abduction is accomplished by increasing its reliability in a hypothesis with additional facts explained, it is a task of cascaded inference, where an explanation found at one level of knowledge is passed along to a more abstract level. For example, findings are explained by a higher level of hypotheses. The overall computation deals with multiple-level processing, in which each level is associated with a specific type of hypothesis and is a cooperative computation with other levels. The number of levels of process to be included in building recognition needs to be determined. Foster (1992) points out that according to research on cognitive psychology, there is a good reason to have many levels based on the context and the purpose of explanation.

Since building recognition is viewed as abductive recognition, or hypothesis-based recognition, knowing the structure of the building parts (e.g. roofs that are likely to be part of the building) and given the measurements that can be made on the image, we can generate and test hypotheses to describe the building components. The proposed building recognition is based on both data driven to generate primitive features and hypotheses driven to explain those primitive features. For the data driven, when we look at our environment, the visual input information is projected onto the retinas of our eyes and at least five stimuli could be involved: (1) luminance differences, (2) color differences, (3) texture differences, (4) binocular disparity, and (5) relative movement between parts of the image. To process the above five stimuli, four assumptions are generally used to guide the research on low-level vision: (1) early vision computation is highly parallel and local, (2) early visual computation has a functional organization, (3) the physics of image formation normally ensures certain correlations between the
properties of the world and the image, and (4) the output of early visual processing is about what primitives the early vision computes (Kalat, 1992; Gregory and Colman, 1995). For the hypotheses driven, we can relatively easily and rapidly recognize an object if we can identify invariance of the object properties in a scene (e.g. size, orientation, or color); otherwise, we need put various feature information together and make coherent hypotheses about the perceived object. In this research, top-down processing and abduction are proposed.

5.5.3.2 Decomposing The Task

Since space is used to describe things related to size, area, and position, human spatial cognition appears to operate differently in small scale space and in large scale space. Although some fundamental spatial concepts may be applied to both kinds of space, the relative salience of the concepts may be quite different. However, space is fundamental to human existence and has a great influence on human thinking. The integration of both the different scales of object shapes and depth information helps to recognize buildings from aerial images.

The approach presented in this research is divided into two phases of process. The first phase of process is to reconstruct coarse surface and to detect candidates of buildings (e.g. a set of 3-D object-centered volumetric objects). The second phase of process is to reconstruct fine surface at each potential building location, to detect primitive features (e.g. 3-D surface patches and facets), and to give each 3-D primitive
feature an explanation. Then, based on the best explanation of selecting surface facets, the generic building model (see Figure 5.3) will be reconstructed.

5.5.3.2.1 The First Phase of Process — Focus on Attention

The main purpose of the first phase of process is to facilitate the performance by improving the pick-up of information in the visual cues. The major process is to reconstruct coarse surface and to formulate a hypothesis on the possible building location for the next phase of process.

To reconstruct surface from a pair of images, two steps can be taken (Schenk and Toth, 1991; Al-Tahir, 1996). First, stereo matching techniques are used to generate a set of 3-D discrete points such as corner points and edge points by matching image intensity and/or extracted features on a pair of images and measuring their parallaxes. For example, at coarse scale, images match the edges to give a rough correspondence between the two images and then this correspondence is refined by using successively fine edges or intensity values. Second, the set of dense 3-D discrete points are used to reconstruct the surface by using interpolation techniques. The reconstructed surface is called a digital terrain model (DTM) in mapping science or a 2.5-D map in computer vision.

After the surface is reconstructed, two and three dimensional primitive features are collected from raw images and DTM data by using image enhancement, segmentation, grouping, and matching techniques. The DTM data is not just an array of depth information, but has a unique property related to building appearance, namely form (or volume), in an outdoor scene. From the DTM data humps can be detected and
presented as 2-D curves or closed polygons. However, when an object has less intensity contrasting with its surroundings, its edges will be difficult to detected from aerial images using current segmentation techniques. Therefore, in homogeneous areas, the edges of buildings are missed and there will be no humps when the surface is reconstructed.

Nevertheless, for general purposes, aerial photographs are usually taken within two hours of local noon. In this way, shadows will be minimal, and the shadow information may be helpful in recovering the missing humps. For example, as shown in Figure 5.6, first we can hypothesize that darker regions in raw images will include potential shadow regions although darker regions may include other features such as trees, grass, and water. Second, both detected humps and darker regions are incorporated and the maximum vote of sun direction is selected by computing each possible sun direction. Third, if a darker region near the detected humps is satisfied with the relation to sun-hump-shadow, the darker region will be labeled as a shadow region; if not, it will be labeled as a tree, a grassy area, or a water region based on its height information in the DTM data. Then, the rest of darker regions, without the relation to humps, are used to predicate the missing location of the humps. Moreover, each predicated location is incorporated with the straight line(s) found and is counted as a potential building location. The outputs of the first phase of process are a set of potential building locations that will be used as input data for the second phase of process.
5.5.3.2.2 The Second Phase of Process — Details of Reconstruction and Recognition

The goal of the second phase of process is to give the best explanation of a reconstructed building appearance based on the findings of primitive features (e.g. surface patches and facets). Based on surface reconstruction and surface segmentation, the second phase of process, a computational strategy for surface-based recognition, is shown in Figure 5.7.

Beginning with one of the potential building locations in fine scale images, the process of surface-based building recognition can be divided into following steps. First, 2-D corner and edge points are segmented from a pair of stereo images. Then, these points are matched in 3-D by using epipolar, contrast gradient, and disparity gradient constraints. Next, the 3-D corner and edge points are used to reconstruct surfaces. After this the reconstructed surfaces are segmented into surface patches with depth and
orientation discontinuities. Then the surface patches are labeled by such properties as area size, location, elevation, orientation, and surface type. Finally, according to the type of discontinuity, adjacent surfaces are grouped into facets, and the hypotheses of surface facets are parts of buildings, roads, vehicles, vegetation, or unknown objects (see Figure 5.3). Moreover, if desired, the surface facets can be approximated by using surface model (library) to reconstruct the surfaces. Subsequently, the next potential building location is sought and the above six steps are repeated until all the potential building locations are visited and explained.

Prior knowledge (task dependent such as image formation, site location, and association)

indexing and matching ↑↓

Surface facets (surface properties such as surface types, size, location, orientation, shape, and brightness)

Surface patches (surface properties)

grouping ↑

smoothing & segmentation ↑

Refine surface

interpolation ↑

Raw surface

matching ↑

Hierarchical stereo

Figure 5.7. The Second Phase of Process for Surface-based Recognition
5.5.4 A Note on Surface Reconstruction and Segmentation

Surface reconstruction is one of early vision processes that do not require domain-dependent knowledge, but fundamental assumptions about physical world (Poggio et al., 1985). Various techniques for surface reconstruction have been developed from stereo matching (Schenk et al., 1990; Cochran and Medioni, 1992) or active range sensors (Besl and Jain, 1986; Li, 1992). A detailed demonstration of surface reconstruction from a pair of aerial images can be found in Al-Tahir (1996). And one of the results of surface reconstruction is DTM data, a 2.5-D image of the 3-D world. The use of DTM can facilitate the measurement, description, and recognition of object shapes based on their surface geometric characteristics. Therefore, 3-D reconstruction of visible surfaces continues to be an important research topic in CV and DP communities.

In surface segmentation, at least three kinds of surface boundaries (edges) need to be detected: jump edge, crease edge, and smooth edge. Jump edges refer to depth discontinuity; crease edges correspond to surface creases, the points over which surface normals are discontinuous; smooth edges are characterized by the continuity of surface normals but the discontinuity of curvature (Hoffman and Jain, 1987). It is known well in differential geometry that the six coefficients of the first and second fundamental forms uniquely determine a smooth surface in Euclidean space. Moreover, the coefficients of the first and second fundamental forms are dependent on the choice of the parameterization. Instead of parameterization, Besl and Jain (1986), Ponce and Brady (1985), Fan et al. (1986), and Li (1992), used curvatures (e.g. principal, mean, and Gaussian curvatures) to classify surfaces. However, the problem of detecting, localizing,
and describing the discontinuities (or edges) in DTM or in 2-D intensity image is to
develop rich representations of shapes. Besl and Jain (1985) summarized 3-D
segmentation techniques in their excellent survey; in addition, an experimental
comparison of range image segmentation algorithms can be found in Hoover et al.
(1996).
In this Chapter, a few experimental results were illustrated. The images used in the experiments were either synthetic or real images. In the following examples, the proposed two-phase processes were demonstrated.

6.1 An Example of the First Phase Process

The first example showed the first phase process, intended to find potential building locations in aerial image. A pair of aerial images, named SUBURB, along with one DTM, were made available to the ISPRS WG III community (Fritsch et al., 1994). The size of the original aerial images is 1K x 1K and DTM resolution is 1 m x 1 m (the size is 238 x 237). For the purpose of studying, both data were resampled to the same size (e.g. 250 x 250).

Based on the proposed first phase process, detecting humps and hypothesizing shadow information were combined to generate potential locations of buildings. The results were presented in Figures 6.1.
Figure 6.1 (a) SUBURB right image (size 250x250)

Figure 6.1 (b) DTM data of SUBURB images, (size 250x250, resolution approx. 1m x 1m) display on 256 gray values

(Continued)
Figure 6.1 (c) Detect humps on DTM data

Figure 6.1 (d) Select Humps as Potential Buildings

(Continued)
Figure 6.1 (e) Potential building locations (mark with circle)

Figure 6.1 (f) Anisotropic smoothing SUBURB right image
(image size: 250 pixel x 250 pixel)

(Continued)
Figure 6.1 (g) Hypothesize darker areas as potential shadow areas (gray value less than 30)

Figure 6.1 (h) Compute possible sun angle from selecting humps and the darker areas
Figure 6.1 (i) Potential shadow locations (satisfied with sun-hump-shadow relationship)

Figure 6.1 (j) Potential shadow locations without selecting humps (Continued)
Figure 6.1 (k) Predicated locations of buildings
(mark with squares)

Figure 6.1 (l) Potential locations of buildings
(mark with circles and stars)
6.2 Examples of the Second Phase Process

The second example was a 100 x 100 synthetic mountain-roof image. All the surfaces were developed, as shown in Figure 6.2 (a). Figure 6.2 (b) showed 3-D view of mesh plot of synthetic image. The result of segmentation of synthetic image was shown in Figure 6.2 (c). The superimposed segmentation result and synthetic image were shown in Figure 6.2 (d).

The third example was a 281 x 308 range image obtained from the web site, http://marathon.csee.usf.edu/range/seg-comp/images.html, the University of South Florida. The process of surface segmentation was similar to the previous example and the results were shown in Figure 6.3.

In the fourth example, the second phase process was demonstrated. The left and right images of the original SUBURB images were resampled to the size of 500 x 500 pixels. Then, a portion of the images was selected from the result of the first phase process and the images included a composite house and other features (e.g. tree, road, grassy area, ditch, and shadow). To begin the second phase process, we assumed that we were in a fine level and we had a coarse DTM data (e.g. 1m x 1m resolution). The second phase process included two steps. The first step involved the details of surface reconstruction, in which the coarse DTM and fine resolution of 3-D discrete matched points were combined to interpolate fine resolution DTM. In the second step, the fine resolution DTM was segmented into surface patches and each surface patch was labeled with a meaning. The results of the second phase process were shown in Figure 6.4.
Figure 6.2. Segmentation of a Synthetic Mountain-roof Image
Figure 6.3. Segmentation of a Test Range Image
Figure 6.4. The Results of the Second Phase Process

(Continued)
Figure 6.4 Continued

Figure 6.4. continuous
Figure 6.4 continued
Figure 6.4. continuous
6.3 Discussion

In the first example, the final result was quite impressive because most buildings in the image were detected. However, two weakness were shown in the first phase process. First, the potential locations of buildings included some false locations of buildings. Therefore, there is a need to generate some new hypotheses to eliminate the false locations of buildings such as checking straight line(s) or parallel lines at each potential location of buildings. Second, the thresholding (hard constraint) of the value of the potential shadow intensity cannot cover all locations of “true” shadows. Therefore, combining other sources (e.g. color images) is helpful in feature classification such as distinguishing between shadow and grass.

The second and third examples showed that although differential geometric quantities have nice invariant properties, they require that data have smooth surfaces.

In the fourth example, from the results we can realize that segmentation, stereo matching, and interpolation all need to improve. For example, a more efficient and effective method to find correspondence between two images is needed because stereo matching plays a very important role in surface reconstruction. In addition, segmentation is important especially in finding corners and crease edges on roofs. And flexible interpolating values in homogeneous areas are needed. Furthermore, combining depth information from stereo matching and active sensors (e.g. range data) is useful in reconstructing surfaces. For surface segmentation, although surface types and normals were computed based on fine DTM, the segmentation result was poor at locating the boundaries of surface patches. The reason may be that with the quantization and noise
errors, the boundaries of the surface patches on DTM are unclear. Nevertheless, using height constraint to detect jump edges, we can locate most part of roofs. However, more researches on 3-D segmentation are needed.
CHAPTER 7

CONCLUSION AND FUTURE RESEARCH

The primary goal of this research is to propose a system for the representation of building recognition knowledge. The challenges of this task include establishing a generic building model (representation), effectively searching a potential building location, and hypothesizing a set of 3-D primitive features (recognition) in a scene. Although this is a highly restricted problem, it is very important and is frequently required in photogrammetric mapping.

It is believed that the ability to apply many constraints to an image in a local and globally consistent manner requires a complex information process. With this process, we can find proper constraints so that the desired interpretation can be achieved. In this research, surface patches and facets are adopted as primitives to recognize buildings in a scene. Reasoning and process strategies are suggested to generate surface patches, facets and their properties. Moreover, prior knowledge for building recognition such as depth and form are used to reduce the search space since aerial images are highly complex and it is difficult to explore different stimuli in the images. For example, a small amount of
prior knowledge (e.g. image formation, site location, and building structure) is brought to the building recognition task by using abduction. In this case, flat or curved surfaces are evidence of roofs of buildings. Although in practice automatic scene recognition of aerial images still has a long way to go, with this research we are encouraged to contribute to the task at hand.

Further research in building recognition should focus on the following areas. First, the data obtained from multi-sensors may be combined together to obtain more reliable information because some sensors may provide specific information to reduce the amount of ambiguity in measuring spatial properties. In this case, an airborne laser scanner can provide range data useful in interpreting objects at homogeneous intensity regions. Second, the introduction of a reliable factor, such as the weight associated with a hypothesis, would allow us to take into account the possibility that a hypothesis correctly or incorrectly matches primitive features. Third, domain-dependent knowledge is very useful for high level of processes in machine vision. It not only simplifies the representation of the task but also reduces the search space between the model and the image primitive features. Therefore, for scene recognition both part/whole and geometric/spatial relations between objects in a scene need to be more clearly described. Fourth, the problems of machine vision such as segmentation, grouping, and matching are vital and should be further investigated. Furthermore, any techniques that can provide more accurate information will have a significant impact on the use of the new data.


98


