USING PERCEPTUAL GROUPING TO DETECT OBJECTS IN AERIAL SCENES [1]

Keith Price and Andres Huertas
Institute for Robotics and Intelligent Systems
University of Southern California
Los Angeles, California 90089-0273

Abstract

Mapping, cartography, photointerpretation and guidance are some applications that can directly and readily benefit from automated aerial scene analysis. We have successfully used perceptual organization ideas to analyze aerial scenes and to describe cultural features of interest, such as buildings. Perceptual organization refers to the ability of a visual system to quickly capture representations of structure and similarity among otherwise random elements, features, and patterns in the visual field. We describe some of the systems we have developed and applied to aerial images and present several examples. We also give a brief description of our current work, in particular the development of the concept of a *grouping field* to represent salinecy and help define the computational aspects of grouping operations. This paper is based on the work of many different people in our research group with more details of each part found in the referenced papers.

KEY WORDS: Automatic Mapping and Photointerpretation, Aerial Image Understanding, Perceptual Grouping, Computer Vision.

1 Introduction

Perceptual organization refers to the ability of a visual system to quickly capture visual representations of structure and similarity among otherwise random elements, features, and patterns. These representations are the result of grouping operations that give the system or individual a sense of the objects in the visual field. There has been a great deal of interest and research on the phenomena of perceptual organization. Originally, perceptual grouping was studied by Gestalt psychologists in the 1920s and 30s. Unfortunately, while they provided many useful insights into the problem and many compelling demonstrations, they did not provide a computational theory.

Much of the work on a computational theory in the image understanding domain has focused on grouping of dots and lines [5, 14, 32, 37, 39, 59, 61, 63, 44]. Perceptual organization ideas are difficult to formalize and several authors including many in our group at USC, are working on the computational aspects of perceptual groupings. Informal derivations are hard to implement in computer vision systems due to the difficulty in detecting grouping relationships, due to our lack of understanding of suitable representations, and the processes that can make use of the established relationships in higher levels of perception [44].

One of the poineering efforts on grouping features in real scenes was done in our group [44, 46, 47]. This system has been applied to building detection as well as object level segmentation from monocular images of complex indoor scenes. We do not claim that this is a computational theory for perceptual grouping and are not aware of such theory in the literature. However we can claim that the use of perceptual organization principles is perhaps almost a must for aerial image analysis, and in particular, in the detection and description of cultural features. There are two main reasons for this. First, perceptual organization makes explicit the geometric relationships among *perceived features* (abstract internal representations of actual features in the visual field). Second, it provides *focus of attention*, that is, a collection of techniques designed to draw attention to significant structures in the image. Note that these two major visual abilities are applied recursively at all levels of perception. We should also point out that at the lower levels of vision we prefer to think of groupings for non-purposive perception (geometric structure without functional attributes) while at the higher levels we like to think of groupings for purposive perception (geometric structure with functional attributes).

In this paper we discuss our more recent work on 2-D and 3-D analysis of aerial images of real scenes. The 2-D work deals mostly with shape issues and is applied to monocular images, and the 3-D analysis deals with objects descriptions in 3-D using either shadows of stereo to obtain 3-D information. Our relevant past work, which applies mostly to the detection and description of cultural features is given in [26, 27, 30, 47, 44].

In the following sections we first give an informal discussion on perceptual organization and describe related work by our group and by other researchers. Next we describe our systems, and last we give a brief description of our current work. The reader is referred to the surveys of Lowe [37] and Palmer [52] for further references on perceptual organization, and to the various references describing our work.

2 Perceptual Organization Issues

Our aerial image understanding work has concentrated on the analysis of cultural features. These features represent structures that are not random but have specific geometric properties. On sunny days and under favorable imaging conditions 3-D structures cast shadows and exhibit other features that allow inference of the 3-D structure from 2-D images. We have used techniques of *perceptual grouping*, *shadow analysis* and *shape from contour* for utilizing these observations. In this section we discuss some of the relevant issues of perceptual grouping. Shadow analysis is briefly discussed later in connection with our early system to detect building in monocular images. Shape from contour is beyond the scope of this paper.

[1] This research was supported by the Defense Advanced Research Projects Agency under contract F49620-90-C-0078, monitored by the Air Force Office of Scientific Research. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

2.1 Perceptual Grouping

Humans, when presented with a set of simple features such as dots or lines, are able to group them into meaningful structures immediately. This process is usually referred to as *perceptual grouping*, and we argue that it is of crucial importance in the process of aerial image analysis. This is due to the large amount of detail in the images and the complexity of the structures in them. Perceptual grouping allows us to separate the meaningful features from the others.

In general terms we can think of the problem of perceptual grouping as consisting of three related subproblems. The first is to determine how to represent the visual features. The second is to determine what kinds of grouping operations to apply to obtain meaningful groups. The third is that there are many possible groupings and we must be able to choose among them.

The human process of perceptual grouping is not fully understood but the grouping criteria are believed to include proximity, collinearity and symmetry of the features and formation of closed figures. We believe these criteria to be helpful for machine processing of aerial scenes. However, in this task we are further aided by the knowledge of the kinds of structures we are interested in. For example, for elongated features such as runways, we expect to find anti-parallel pairs of lines (itself a simple symmetric grouping); for buildings we expect to find regular geometric structures.

Points and curves are useful representations. Points denote position but can have other attributes such as size that denotes extent of the feature. Straight curves (i.e. line segments) denote visual boundaries as well as directionality. They can also denote symmetries.

For selection among various possible groupings, we can use multiple criteria. For example, if a curve forms symmetric pairs with two other curves, we can choose the one that gives a more closed or compact figure. Essentially, we can allow the groupings to compete and cooperate depending on whether they are mutually exclusive or supportive. As we will see below, we have used such a scheme in the past to successfully detect groups of line segments that might correspond to roofs of buildings in aerial images [47] and to segment scenes of complex man-made objects without specific knowledge of the objects in the scene [44].

Another example of perceptual grouping is in detecting roads, runways and taxiways in aerial images. Here, the process of grouping starts with finding anti-parallel pairs of lines and grouping the pairs that are collinear. Only those groups that are long and meet the other criteria are preserved. (For runways, the test is to look for specific surface markings). Our previous experience with perceptual groupings is described more fully in [26, 27, 30, 44, 47].

Most work in the IU community on object recognition uses a "model-based" approach. We believe that much of this work really should be viewed as solving the pose estimation problem rather than the object recognition problem. One deficiency of the methods is that they require very specific shape models. For example, it is not sufficient to say that the building is a rectangular parallelepiped; we must also supply the relative dimensions of the sides.

ACRONYM, a model-based system developed in the early 1980s, attempted to use more generic models of objects such as airplanes [6, 7]. However, the descriptive abilities of this system were highly limited and the recognition process consisted mostly of checking whether certain kinds of "ribbons" were present in the image. At USC, we have an on-going effort in description of complex shapes modeled as "Generalized Cylinders" [9]. Similar research, though using a somewhat different approach, is being conducted at Stanford University [57, 56].

2.2 Our Approach to Perceptual Grouping

We believe that the grouping operations that yield organized perception consist mainly of symbolic evaluation of the properties of the features that are candidates for grouping. By features we mean representations of visual elements; by grouped features or feature groupings we mean geometrically and structurally significant groupings along a hierarchy of recursively formed groups, from points to complex objects. By object detection we mean two things: the description of the shape of the objects and the description of their structure. The shape descriptions should be at various scales using features that are invariant to changes in the viewpoint.

In the following sections we summarize our approach to defining and constructing meaningful feature groupings. For additional details see [44].

2.2.1 Similarity Issues

It appears that humans prefer to group elements that have similar characteristics such as shape, intensity and color [52, 60]. We do not explicitly use color or intensity but, as mentioned above, we use two primitives, points and curves, to represent position and shape, and thus the similarity is among these simple features.

2.2.2 Structure and Scale Issues

The relationships among the points and lines convey structure and are thus a criteria for grouping. Given the importance of structural information in visual processing [61], this has been the most studied component of perceptual organization in computer vision and psychology [37, 39, 46, 47, 52, 54, 59, 63]. Some grouping processes use only structural relationships, and others are hierarchical. We can separate the elements into distinct groupings even though they do not correspond to any objects we recognize.

Structural groupings can be further subdivided on the basis of scale. At small scales the structural relationships form locally, at the level of subparts or parts of objects such as in regular textures (see [63]). Although the structural grouping process outlined in [44] can detect such groupings as well, in the work described here we concentrate on groupings at the scale of the objects in a scene such as buildings and runways.

2.3 What to Group?

The Gestalt psychologists believed in the principle of Prägnanz (goodness or simplicity of form) as a fundamental criterion to group elements. Recent work has tried to develop computational criteria leading to explanations that are primarily geometrical, for example in terms of transformational invariances [52], or probability [37, 61], or in terms of mechanisms such as orientation selection [63]. For groupings at the level of objects we have preferred a more functional explanation based on the *significance* (identification of structures in a image that have high likelihood of corresponding to object structures, i.e. focus of attention), *representation* (usefulness to other visual processes) of grouped features, and *selection* among multiple groupings (greater saliency).

2.3.1 Significance Issues

Our interest has been in non-natural objects, most of which exhibit a great deal of geometric regularity. The principle of *non-accidentalness* [37, 61] states that "regular geometric relationships are so unlikely to arise by accident that when detected, they almost certainly reflect some underlying causal relationship." The probability of two lines, which are not parallel in 3-D, projecting in the image as parallel lines in 2-D (due to an accidental alignment), can be considered so small that we can with high confidence claim that parallel lines detected in 2-D are also parallel lines in 3-D.

We have designed our systems to favor and be sensitive to the shapes and structure of the objects they were designed to detect. We detect viewpoint-invariant structural relationships that are common in the objects of interest and use the non-accidentalness principle to reason that the detected groups were caused by the structure of the objects in the scene. We choose groups that identify structural arrangements of visual elements that have a high probability of corresponding to object of interest.

2.3.2 Representation Issues

Grouped elements must encode the geometric and structural criteria that led to its formation. They basically represent higher levels of abstraction but must allow recursive recovery down to the primitive image elements. The chosen representations for groups should clearly allow the system to perform segmentation, description, focus of attention and integration of evidence at all levels.

- Segmentation: Relationships among substructures may also exist among structures and superstructures.
- Description: Most object shapes are described in terms of component shapes. The hierarchical decompositions should be easily identified.
- Focus of Attention: By encoding appropriate and relevant

grouping information, the groups should provide guidance and contextual cues.

- Integration: As much as possible groups represent relationships that are invariant to viewpoint, and thus can be used to integrate information from multiple views, site models, multiple sensors and multiple media. In correspondence processes such as stereo, motion, and model matching, improved performance has been obtained by using more abstract features [43, 45, 36, 18]. This also results in a significant reduction in the computational expense of matching. Perceptual groups are more complex than edges, thus there is much less ambiguity in matching higher level features, as we show in examples later.

2.3.3 Selection Issues

The inherent complexity of aerial scenes in terms of detail and in terms of the possible combinations leads to a large number of possible groupings. In practical terms a computer vision system must include a set of parameters that keep the data bases within a reasonable size. On the other hand there is the issue of fragmented and incomplete low level information due to image content, quality, resolution, etc.

The systems described by Huertas et al. [26, 27, 30] deal with fragmented information and generate minimal groupings at the expense of generality. In the systems described by Mohan and Nevatia [44, 46, 47], all reasonable groups are formed to deal with fragmentation and incompleteness at the expense of requiring a selection mechanism. They decided on the "goodness" of a group on the basis of how it compares to its alternatives in terms of the support it has from related groups at other levels, and the support or contradiction from its component primitive features and other related image features. Furthermore a given group representing a level of abstraction is supported not only by its component subgroups but by the supergroups it is a component of. In general terms, groups which are linked by part-of relationships are mutually *supportive* and those that share component subgroups are mutually *competitive.*

The problem of selecting the best set of groups in [30, 27] is formulated as that of selecting those hypotheses that could be verified as a representation of an object of interest: matching shadows for buildings and surface markings for runways. In [44] the problem is formulated as that of selecting the best set of hypotheses, given relationships of support and conflict among them. To choose the best consistent set of hypothesis that maximize evidence, we wish to assign confidence values to hypotheses such that the cumulative value of the support, the conflict and the contribution of all the underlying evidence is maximized. Our goal is to find the *optimal* feature groupings consistent with the known optical and geometrical constraints [4, 11]. Note that all the constraints must be *simultaneously* satisfied to reach global consistency across all levels of the hierarchy.

One parallel technique to solve this problem is relaxation where a cost function associated with the network is minimized. We wish to select the best *consistent* feature groupings, and reject the bad groupings. If we formulate the cost function such that the optimal solution corresponds to its global minima, then the problem of locating the best groupings is reduced to optimizing the cost of the network given the constraints between the grouped features and the observed image characteristics. Parallel optimization techniques such as simulated annealing [33], Hopfield networks [24], Boltzman machines [11, 12, 55], probabilistic solutions [19] and connectionist methods [13, 55], have been proposed for such problems. In our system [44] the Hopfield formulation was used to implement the constraint satisfaction network.

2.4 Related Work on Perceptual Grouping

Mohan [44] provides a discussion on the relationships and differences between our approach to perceptual grouping and that of others. Here we paraphrase some of his discussion.

2.4.1 In Psychology

Lowe [37] presents an excellent survey of the work, both in psychology and computer vision, on perceptual organization techniques. Palmer [52] also provides a detailed survey of work in psychology pertaining to perceptual organization.

To characterize human preference for some shapes and arrangements over others the Gestalt psychologists believed in the principle of "simplicity of form." However, they did not formalize this notion in computational terms. The notion of "goodness" or simplicity was later quantified by other in terms of information theory and coding theory (the simplest form was one encoded in the minimum number of bits) [1, 23, 35].

Garner and Clement [16, 17] explained observed goodness ratings for simple geometric patterns in terms of their invariance to certain transformations. Palmer [52] develops on their work. We follow Palmer's theory, namely that perceptual organization detects those geometrical relationships that are invariant to viewpoint-transformations.

2.4.2 In Computer Vision

Marr [40] viewed perceptual organization as grouping processes that operate on the "raw primal sketch" (primitive descriptions of the image) to build up descriptive primitives. Furthermore, these descriptors are built in a recursive manner with, at each higher level, the primitives referring to increasingly abstract properties of the image. In these two basic points, that perceptual organization is a process to generate descriptors, and that the descriptive hierarchy so formed is recursively generated to describe more and more abstract properties of the image, Mohan's thesis and Marr's are the same. However, there are substantial differences on almost all other points: differences in scope and implementation, in the choice of relations captured by the grouping, and in the causes behind these choices.

A paper by Witkin and Tenenbaum [61] on the role of structure in vision includes three broad observations that we believe are relevant to the subject of this paper:

- Inadequacy of local and intensity based techniques,
- The concept of non-accidentalness, and
- The use of perceptual groupings as intermediate descriptors,

and we will discuss them in that order.

They note that one major trend in computer vision focused on recovery of local quantitative surface properties, such as depth and orientation, and point out that such techniques do not perform well. They attribute the poor performance of these techniques to the local, quantitative approach. Further, they present various examples where we are easily able to detect structure in images, and where the simple models used by these local techniques break down: in complex images where the reflectance functions of the surfaces change rapidly and for which we do not know the camera parameters or the source of light, in images where intensity encodes some other imaging characteristic such as in range images or electron micrographs, in drawings where intensity information is missing, and even in images constructed such that intensity patterns contradict the structure. Even if arrays of surface properties are recoverable in a reliable manner, they note that using them is not a straightforward task. These observations suggest that alternative, qualitative, modes of obtaining (and representing) structural information, rather than the quantitative ones, should be explored.

Witkin and Tenenbaum state the non-accidental criteria as: " ... regular relationships are so unlikely to arise by chance that, when such relationships can be consistently postulated, they almost certainly reflect some underlying causal relationship and therefore *should* be postulated. We conjecture that as a least-distortion solution approaches strict identity, the likelihood that the relationship is non-accidental increases. The minimization of change is therefore a primary basis for discovering causal relationships at a primitive level." They note a number of specific regular relationships, but they provide neither a specific mechanism to detect such relationships nor do they attempt to identify structural relationships significant for particular visual tasks or domains. In our formulation of perceptual organization, we augment this non-accidentalness criteria with the conditions that the structural relationships considered should be related to the geometric properties of the objects in the visual domain, and that they should be invariant over the set of relevant viewpoints. These conditions not only help us decide on the specific structural relationships to choose for a visual domain (especially when taken in conjunction with the condition of utility, i.e. they should also be useful for other visual process), but also give weight to arguments of causality as we can say that the relationships are not only unlikely to arise by accident but also correspond to objects since they are obtained from object geometries.

The way Mohan's select the significant groupings differs from that proposed by Witkin and Tenenbaum. They suggest using some measure of distortion along least-variational principles (no specific examples are provided) and suggest assigning significance to a grouping based on prior probabilities of observing a relationship with the given distortion. However, they accept that there is no suitable way of calculating these probabilities for real images. Mohan's selection process is also based on a specific measure for each grouping (the measure can be interpreted as a quantification of distortion) but the selection process is based on comparison to alternates, and relationships to supporting groupings, rather than a direct assignment of probabilities (followed possibly by some threshold on the probabilities). It seems Witkin and Tenenbaum do not consider that there could be various alternate groupings for the same features. On the other hand, in our experience with real images, one basic problem is the multiplicity of possible groupings, which we have to restrict to manageable numbers by use of both heuristics and selection techniques.

While Witkin and Tenenbaum do not state explicitly that the primary task of perceptual organization is to generate an abstract feature hierarchy, it is clear that they view groupings as descriptors: edges, coherent regions, groupings, flow patterns, parallelism, symmetry etc. These entities, that they call "semantic precursors," are "decomposition of the image into discrete chunks each of which reflects some underlying cause or process, and therefore must be explained." However, they do not provide either specific description hierarchies or "explanation" mechanisms for any of their proposed structures. In Mohan's work he has provided specific structures for use as descriptors, and by casting them as features, has shown how they are used in visual processes such as object segmentation, shape description and matching.

2.4.3 Applications

Using the "viewpoint invariance" criterion, which states that in general we can assume that the camera viewpoint is independent of the object arrangement, Lowe and Binford infer various 3-D relationships given certain 2-D relationships in the image. This 3-D from 2-D inference is implemented to detect height (of airplanes) by connecting surface boundaries to their shadows [37, 39]. In more recent work, Lowe has extended the proposal of Witkin and Tenenbaum by computing prior probabilities of accidental and non-accidental instances of certain relations (collinearity, proximity of end-points, and parallelism) and using this to constrain the search in model matching. The prior probabilities are computed assuming a random distribution of straight lines in the scene. The groupings are used primarily to reduce both the number of features for matching and the possible transformations (camera viewpoint and object orientation) suggested by the match.

We believe there are serious shortcomings with the approach of computing prior-probabilities for assigning significance to groupings. In the scenes used by Lowe in his matching system [37, 38], there is either one object in unknown orientation, or a jumbled group of identical objects all composed of straight boundaries. For these scenes, there may be some basis in assuming the image model to be a random collection of straight lines. We agree with Witkin and Tenenbaum that there is no suitable way of assigning prior probabilities to groupings in general for real scenes. For example, in scenes of curved objects, the symmetry relation is not as specific as that of parallelism, and it would be difficult to compute prior probabilities of all possible curve relationships which can be labelled as symmetries for all possible types of curves.

There is another potential problem with computing prior probabilities for real scenes. In cases, such as urban areas, inside office buildings, in factories etc., the scenes consist of an organized layout of objects, and the assumption that the objects are randomly oriented breaks down. For the domain of aerial images of buildings in urban areas, the assumption of random placement is violated in a domain for which we have made successful use of perceptual organization. In fact, as we show later in this section, perceptual organization has been used to exploit this very fact of organized layout of objects in a scene [53, 54]. We believe that the approach used in our work, on basing the significance of a grouping on its comparison to its alternates and its relationships with related groupings in the hierarchy (by part-of relationships), is a more reasonable solution.

Lowe (and others [31] who have employed grouping to constrain matching) do not use the groupings to generate a complete description framework for the objects in the scene, and thus have not suggested their use to either segment objects or to describe them. Lowe proposes that segmentation is one important task of perceptual organization but proposes no mechanism for it. He states that "A major reason why perceptual organization has not been a focus of computer vision research is probably because these groupings often do not lead immediately to a single physical interpretation." In our work we have found that perceptual organization gives us useful high level features. Even if we obtain multiple groupings, or even some wrong groupings, from the grouping process (in contrast to unique features as in edges or regions) it is still simple to perform correspondence and segmentation with the groupings computed in 2-D, or with additional information from stereo, to obtain unambiguous segmentation at object level.

2.4.4 Scene based organization

In contrast to the work described above (and the work presented in this paper), some implementations have used structural relationships at the scene level rather than at the object level.

Reynolds and Beveridge [54] have studied geometrical organizations in aerial imagery. They detect groupings of parallel lines, and proximate orthogonal lines (among other relationships). A preponderance of such organization would indicate the presence of organization at the scene level. Thus they could potentially have been used to differentiate rural areas from urban areas. However, the groupings that they detect indicate organization at the scene level and do not correspond to individual objects, and therefore, these groupings are not useful for segmenting individual objects such as roads and buildings.

Quan et al. [53] have successfully used scene level groupings to detect and estimate motion. They have chosen the domain of a robot moving in a well organized environment (inside a buildings which has numerous aligned lines at wall boundaries, doors, windows and cabinets). The match between global organizations in different images from a motion sequence becomes a trivial task, and retrieving motion parameters from the match is robust due to the large number of individual features used in each grouping.

3 Detection of Cultural Features in Aerial Images

Perceptual grouping has been the basic approach for much of our work on detecting buildings and other structures in aerial images. While a wide variety of techniques have been applied towards this task, a systematic use of perceptual grouping has been lacking. Another observation is that while non-natural objects have rich geometric structure, little use of this structural information was made in the older systems. We grouped contours with some structural guidance such as oriented corners (L-junctions) and depth from shadows in [28, 29, 30]. We discuss this approach further in the next section.

Fua and Hanson [15] segment the scene into regions, find edges lying on region boundaries and then see if there is evidence of geometric structure among these edges to classify the region as a building or similar object. In the VISIONS system [21], region segmentation is the primary technique used and the regions are classified by their shape and spectral properties. SPAM [42] is a map based system which uses region segmentation of aerial imagery.

Most of these systems work on simple scenes, for example rural scenes, where the building roof can be simply segmented (and even identified) from the background on spectral properties. The detected buildings have simple shapes. Only a few systems compute and use depth information. None of the systems generate a description of the buildings at the level of shape descriptions of the different wings.

For the systems mentioned above, the generic feature extraction techniques, namely region segmentation or contour following, are not suited for extracting particular shapes or organizations. If the features being detected have simple geometric properties, it is more straightforward to use specific detection algorithms. The Hough transform [2, 3] is a general mechanism for detecting groupings, but is practical only if the exact shapes (rather than generic descriptions) are known. The MOSAIC system [22] uses oriented junctions to complete fragmented lines. This system also uses height information obtained from stereo and sophisticated geometrical reasoning to hypothesize likely wire frame models of the buildings. The complexity of this system, and its limited performance, are due to the use of simple features (lines and junctions) to perform the detection, stereo matching and reasoning. Recently, application of perceptual grouping to locate features indicating structure has been explored by Reynolds and Beveridge [54]. This systems also employs specific routines to detect various geometric organizations indicative of structure. However, this system has limited use as the groupings are sensitive to the layout of the scene rather than the object shapes, and consequently can not be used to either detect or describe any individual structures (like buildings) in the scene.

4 Monocular Techniques

Photointerpretation tasks require the detection of a particular set of objects in an aerial scene. Some examples are, the detection of structures, such as buildings [15, 22, 30, 41], roadways and storage tanks [25, 34], runways [26, 27] and airplanes [7, 58] in images of airports, and docks and ships [49]. In these scenes, many of the objects have restricted shapes, the viewpoint is often restricted, and many of the object shapes can be characterized as compositions of small sets of basic shapes.

Given this set of basic shapes, we can deduce the pertinent structural relationships for the domain by decomposing the basic shapes into the component structural relationships. As an image is a 2-D projection of a 3-D scene, we need to select those structural relationships (to characterize the shapes) that project invariantly over various viewpoints. If for a domain, the set of viewpoints is restricted, as is the case for most aerial images, we need to consider transformation invariance only for the restricted set of viewpoints.

This set of shapes is not specific in the sense of having a particular model for each object (e.g. the model of a Boeing-747 [6]), rather the object shapes are arbitrary compositions of known, specific, basic shapes. The methodology presented in Mohan's system [44] can, by simple extensions or modifications, be applied to any domain where the set of specific shapes consists of basic shapes. Apart from changes specific to the selected basic shapes, no changes to the methodology itself should be required to deal with other specific shapes.

Thus given a set of specific shapes for objects in scenes, our system allows us to design a feature hierarchy which encodes the structural relationships specific to this set. In [44, 47] we have considered the case of object shapes composed of rectangles. Lines, parallels, U-contours, and rectangles are identified as the pertinent structures these shapes can be decomposed into.

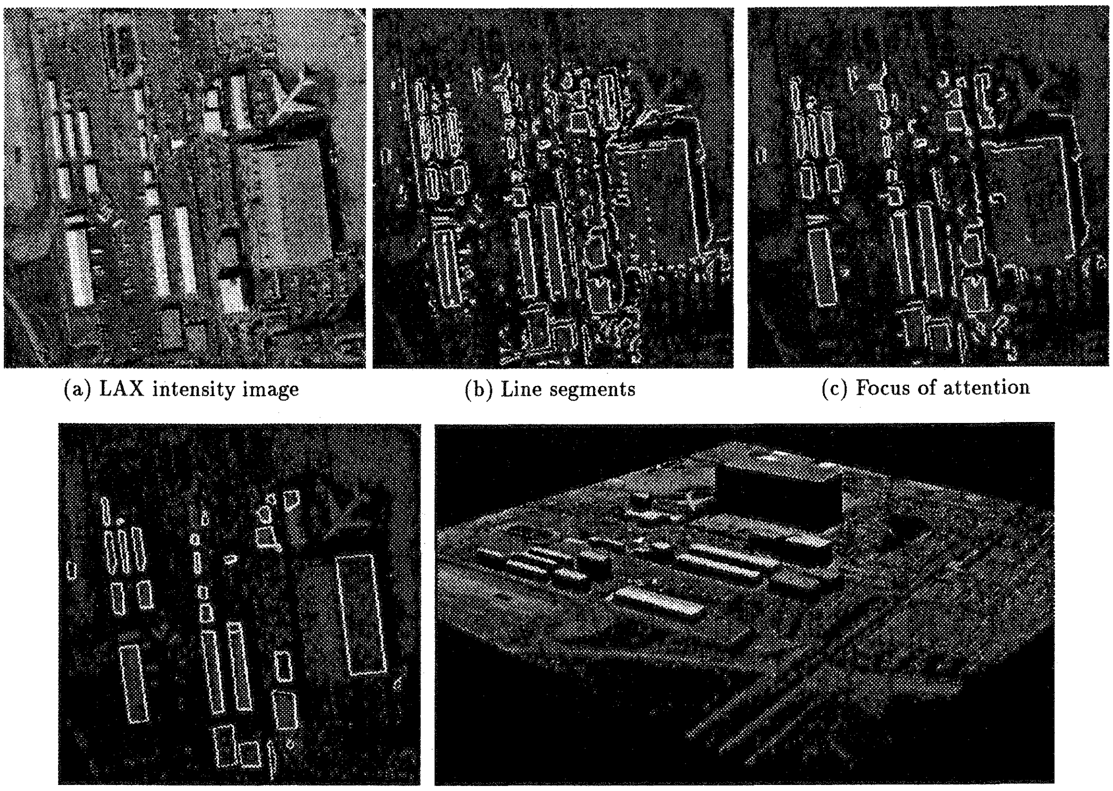
4.1 Detection of Buildings

Our first attempt to group visual features at the level of buildings in aerial scenes is described in [29, 30]. In this system, L- and T-junctions formed by line segments that approximate the intensity edges in the scene provide the *focus of attention*; they are the starting point for a process that follows and groups the object boundaries according to simple geometric constraints. These constraints limit the extracted shapes to rectangles and compositions of rectangles.

The process is aided by initial interpretations given to the L-junctions as possible object and matching shadow corners. As many object boundaries are expected to be fragmented and distorted, nearby partial boundaries are grouped if they have certain geometric configurations and poses with respect to each other.

The shadows cast are the cue to 3-D objects (as opposed to parking lots, swimming pools, and other rectangular structures) and help in estimating the height of buildings. Figure 1a and 1b show an image (from LA International Airport) containing simple rectangular buildings and the line segments extracted from it using our LINEAR feature extraction software [48]. The focus of attention is provided by the line segments forming near 90° L-junctions and T-junctions shown in figure 1c. We make an explicit record of these, and call them "corners". In figures 1d we show the segments grouped into "rectangles", including the missing and hypothesized sides. In figure 1e we show a 3-D rendered representation of the scene including those rectangles for which corroborating shadow information was found. The model of the scene is generated from an arbitrary viewpoint.

A more general illustration of the use of perceptual orga-



(a) LAX intensity image

(b) Line segments

(c) Focus of attention

(d) Rectangles detected

(e) 3D model of scene

Figure 1: Building detection by boundary grouping and shadow processing

nization for the detection of specific shapes is the task of detecting and describing more complex buildings in aerial images. Our method initially detects all reasonable feature groupings. A *constraint satisfaction network* (CSN) is then used to model the complex interactions between the grouped features and select the promising ones.

The buildings are modeled as compositions of rectangular blocks. The roofs of the buildings are thus objects whose shapes are compositions of rectangles. For many overhead viewpoints, the imaging plane is parallel to the ground plane and the roofs. Thus, this real world domain meets both the requirement that the object shapes are composed of rectangles, and the viewpoint is restricted. We believe that our system has allowed us to obtain better results on more complex scenes than the previous systems that employed intensity based features.

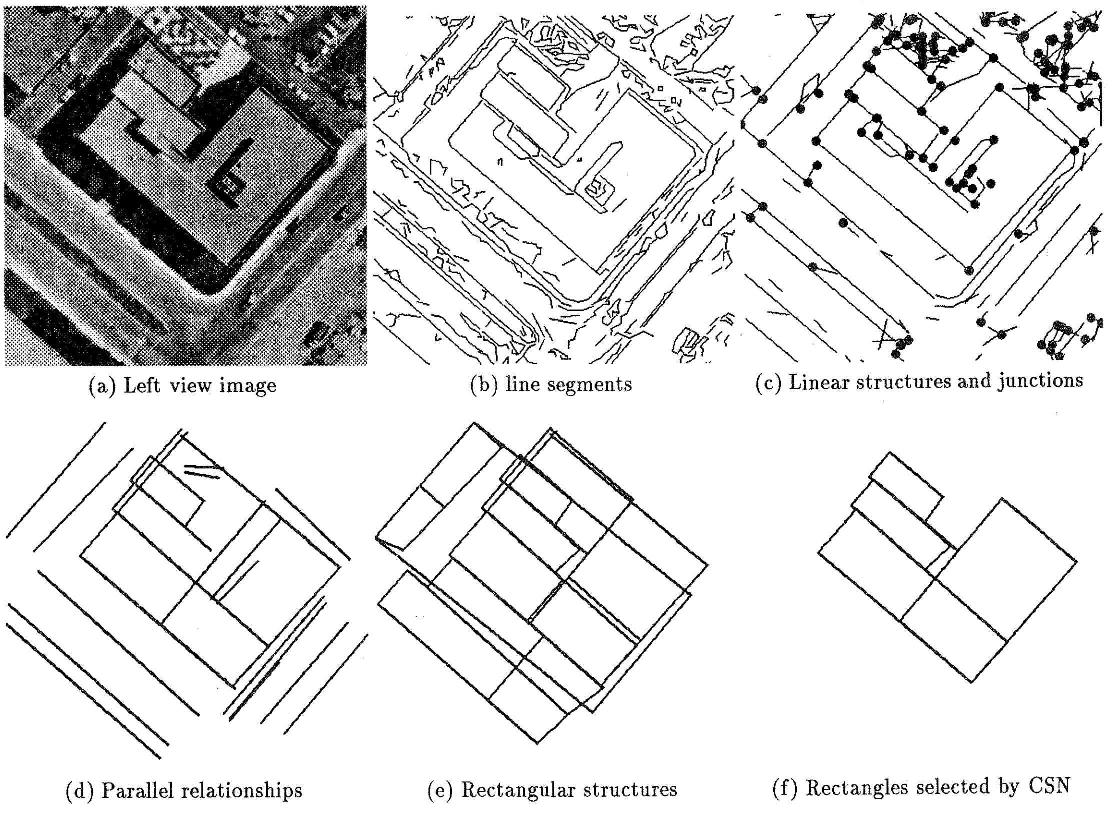
We will illustrate the task of detecting and describing complex buildings in natural scenes by an example. Figure 2a shows the left view in a stereo pair of images of a building in a suburban environment. The building is easy for humans to see and describe, even without stereo, but it is difficult for computer vision systems. Figure 2b shows the line segments detected in the image using LINEAR, the "Nevatia-Babu line finder" [48]. We are still able to see the roof structures of the buildings readily and easily, but the complexity of the task now becomes more apparent. The building boundary is fragmented, there are many gaps and missing segments. There are also many extraneous boundaries caused by other structures in the scene. While local techniques, such as "contour-following" have proved useful for simpler instances of such tasks [30], they are likely to fail for the scene of the complexity shown here.

This task is difficult for several reasons. The contrast between the roof of a building and surrounding structures such as curbs, parking lots, and walkways can be low. The contrast between the roofs of various wings, typically made of the same material, may be even lower. Low contrast alone is likely to cause low-level segmentation to be fragmented. In addition, small structures on the roof and objects, such as cars and trees, adjacent to the building will cause further fragmentation and give rise to "extraneous" boundaries. Roofs may also have markings on them caused by dirt or variations in material. Shadows and other surface markings on the roof cause similar problems.

There are other characteristics of these images which may specifically cause problems for contour following type systems [28, 29, 30]. Roofs have raised borders which sometimes cast shadows on the roof. This results in multiple close parallel edges along the roof boundaries and often these edges are broken and disjoint. At roof corners and at junctions of two roofs, multiple lines meet leading to a number of corners making it difficult to choose a corner for tracking. A roof cast a shadow along its side and often there are objects on the ground such as grass, trees, trucks, pavement, etc., which lead to changes in the contrast along the roof sides. Thus while tracking one can face reversal in edge direction. Often some structures both on the roof and on the ground are so near the roof that the border edges get merged with the edges of these objects, leading contour trackers off the roof onto the ground or inside the roof area. At junctions it is difficult to decide which path to take. Searching all paths at junctions leads to a combinatorial explosion of paths. It may be difficult to decide on the correct contours since contours may not close because of missing edge information, or more than one closed contour may be generated. Contours may merge roofs or roofs and parts of the ground. Figure 2b illustrates some of these problems. Figure 2c shows the lines obtained from grouping the segments in figure 2b.

Structures in urban scenes like building, roads and parking lots are often organized in regular grid-like patterns. These structures are all composed of parallel sides. As a consequence, for each significant line-structure detected in the scene, there is not one but many lines parallel to it. For each line, we find lines that are parallel and satisfy a number of reasonable constraints. These are shown in figure 2d. Figure 2e shows the U-contours and the rectangles formed from the U-contours in this image.

We use a constraint satisfaction network (CSN) to select the few best lines, parallels, U-structures, and rectangles. To insure the selection of *perceptually significant* feature groupings in the scene, the choice of network weights reflects the perceptual importance placed on the optical and geometric constraints between the various grouped features. The perceptual significance


(a) Left view image

(b) line segments

(c) Linear structures and junctions

(d) Parallel relationships

(e) Rectangular structures

(f) Rectangles selected by CSN

Figure 2: Rectangle detection by perceptual grouping

of a grouped feature lies in its indication of actual object structure in the scene. For example, while any grouping of parallel lines [54] is indicative of some order in the scene, we are more interested in parallels that actually correspond to individual objects. Therefore, the parallels that have supporting structural evidence such as rectangles are more significant than those that do not.

While feature groupings at all levels of complexity get selected simultaneously, only the rectangles so selected have been displayed in figure 2f. In our implementation, the weights on the links are not symmetric, so the convergence results for Hopfield networks can not be used. However there is support for that assumption that the networks can converge with non-symmetric weights [8]. We have found that our networks to converge on all selection of weights within ten iterations.

4.2 Airport Analysis

We have pursued the detection of runways, taxiways, the connections among them and aircraft in airports scenes [26, 27] as part of our project to automatically map complex cultural areas. Our long-term goal has been to map all of the interesting objects in the scene and also to devise integrated descriptions that include the functional relationships of the objects in the scene. Runways are complex objects, containing visible signs of heavy use, such as tire tread marks, oil spots, and exhaust fume smears. Runways may be extended or patched using different materials. Runways have a variety of markings (centerlines, sidestripes, threshold markings, distance markings, touchdown marks, etc.) that identify them as such.

Our system first extracts line segments (linear approximation of linked intensity edges). Airport runways are linear features and thus, well characterized by anti-parallel (apars) pairs of segments (having opposing contrast). For an image of a portion of Boston Logan Airport, the line segments and the apars are shown in figures 3a, 3b and 3c. The apars are shown as a line denoting the overlap and the axis of symmetry of the two parallel line segments.

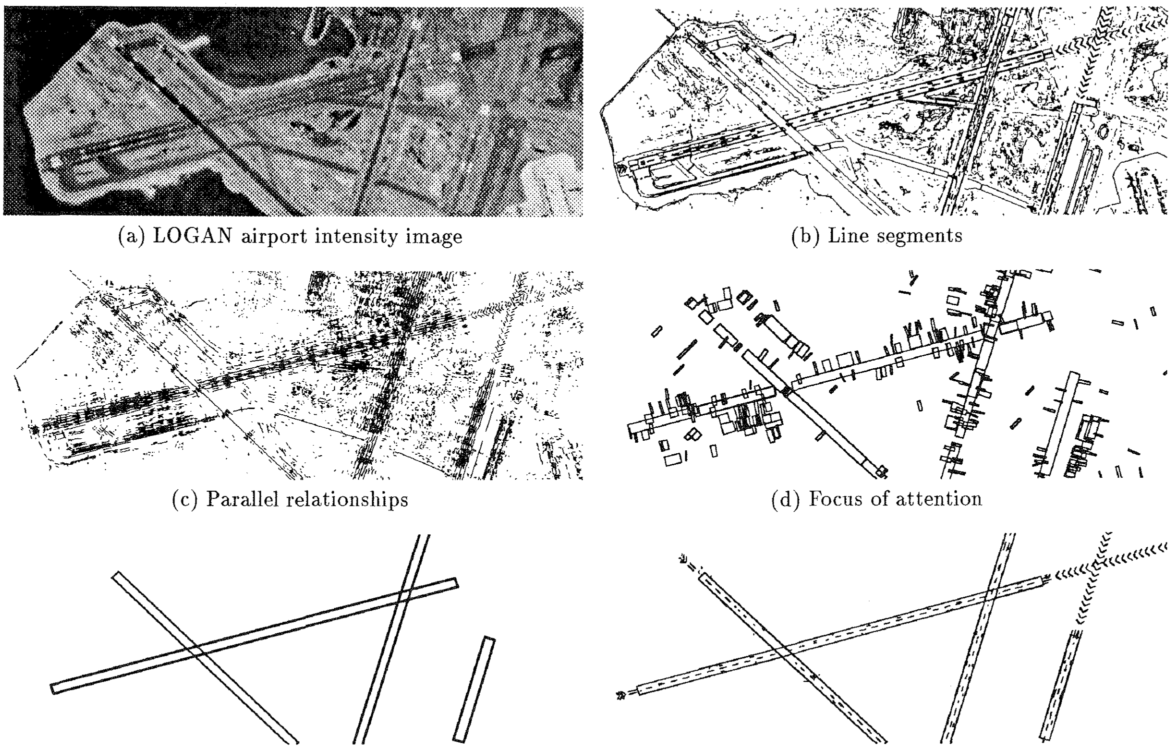
The apars represent low level groupings on parallelism, where the distance between the parallels is a function of the image resolution. All such parallel relationships are computed. The focus of attention is provided by the dominant orientation of the apars and by the apars that contribute to it. The focus of attention mechanism results in over 95% reduction in the search space, and leads to the extraction of the apars representing potential runway fragments (shown as rectangles in figure 3d). Next we proceed to apply continuity and collinearity grouping operations that yield runway hypotheses (figure 3e). Hypotheses are verified by detecting evidence of the markings that we expect runways to have (figure 3f).

A second example is shown in figure 4. The image in figure 4a shows that the runways not necessarily appear as uniform intensity elongated linear structures. Region based approaches may have difficulty in these cases. The perception of linearity however is strong in spite of the irregular shapes of the repair work patches. Runways are extended using different material as the original, they intersect, have different widths, and so on. The line segments extracted from the image are shown in figure 4b. As with the logan example above, and using the same parameters throughout, the dominant orientation of the apars is computed (the peaks in the length weighted histogram of the apar orientations) and used to extract potential runway fragments. These are represented by apars having the dominant orientation and a width consistent with runway design parameters).

The detection of straight portions of taxiways and roadways can be carried out following a similar procedure. For a discussion on taxiways and junctions see [26]. Basically taxiways are much more complex objects than runways, since they have a wider range of geometrical parameters. However, besides generic knowledge we make use of the context provided by the runways to help detect the taxiways. We verify taxiway hypotheses by looking for evidence of markings as well.

4.3 Detection of Pier Areas

We have developed a technique to detect the pier areas in harbor scenes as part of an effort to derive a taxonomy of perceptual grouping operations on primitive visual features like points, lines and junctions. This task is part of our current work to classify grouping operations, and to develop a representation that captures the saliency of a percept in terms of the attributes of a set of features that lead to perceptual grouping. We call this representation a *grouping field* and give examples later.



(a) LOGAN airport intensity image

(b) Line segments

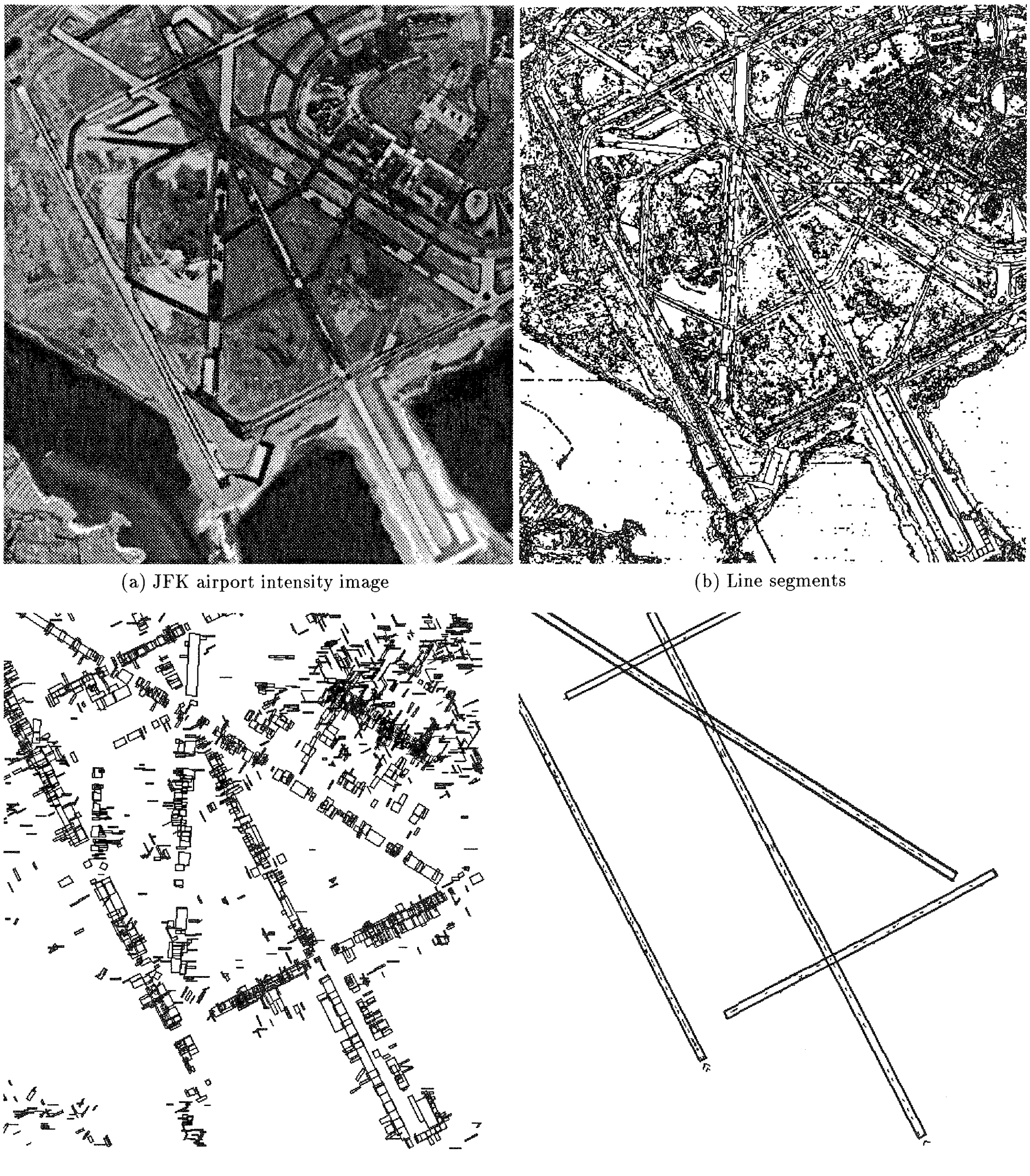
(c) Parallel relationships

(d) Focus of attention

(e) Grouped features

(f) Runways and markings

Figure 3: Runway detection by collinearity grouping



(a) JFK airport intensity image

(b) Line segments

(c) Focus of attention

(d) Runways and markings

Figure 4: JFK Runways detected by collinearity grouping

5 Stereo Techniques

The key problem with general stereo systems is the ambiguity in matching necessitating a mechanism to choose among many competing matches for each match primitive. For our system we have found the constraints imposed by the structure of the grouped features (rectangles) sufficient to select unique matches for the primitives. In the rare case of a rectangle having more than one match, we choose the match with the least number of disparity differences between the sides, which is equivalent to preferring the least occluded interpretation. In the following we summarize from the work of Mohan and Nevatia. Full details can be found in [44, 47].

The rectangles detected monocularly as described in the previous section are used in this system for stereo matching, object segmentation and shape description. Stereo matching is performed on the rectangles to obtain height information. Structural reasoning is performed on the matched rectangles, based on monocular information and 3-D information obtained from stereo, to segment building roofs. The object segmentation automatically provides a shape description of the roof in terms of the component rectangles. The segmented roof area and their heights are used to generate 3-D models of the buildings.

Matching rectangles results in less ambiguity than edges as there are fewer possible alternatives and more information to judge a match. Also there are usually many fewer grouped structures, at any given representation level than edges. The most probable role of grouped features, and one that we employ here, is that correspondence of grouped features provides a rough correspondence for their component primitive features, which can then be matched with less ambiguity. In a similar vein, recent stereo systems have shown improved performance by using more structure than individual edges [10, 36, 43, 45, 51]

For photointerpretation tasks, edge and segment based stereo matching algorithms displayed poor performances. The following factors indicate why stereo systems based on simple image features may not perform well in this domain:

- **Organized nature of the scene.** The are numerous parallel lines since the buildings, roads, parking lots, etc. are all parallel. This leads to the same type of problems as with monocular analysis.
- **Absence of texture.** The buildings sides mark regions with high disparity differences and there are insufficient markings on the roofs to support match-disparities at roof level while matches giving low disparities get favored due to the preponderance of features on the ground.

The choice of rectangle as the match primitive restricts the possible matches. Like other stereo matching systems we allow only matches falling within a disparity range reasonable for the stereo pair. To avoid mistaking rectangles corresponding to tennis courts, parking lots and the like, the legal disparity range should start just above ground level. The other end of the interval should be high enough to encompass the tallest buildings in the scene. This estimate need not be exact, as wrong matches between rectangles usually result in disproportionate disparities. For our test cases we chose an ad hoc value which was more than twice the disparity of the tallest building in any of the test scenes.

Stereo serves as an important visual clue in selecting those grouped features which have a very good chance of corresponding to actual object structures, in this case the roofs. Selection of the proper grouped features is crucial for this domain as many other objects in the scene such as road segments, parking lots and sidewalks have rectangular structures. Furthermore, these objects are arranged in a regular grid like manner, and some grouped features formed reflect the structure in the layout of the scene rather than that of specific objects. The rectangles selected by out system are shown in figure 5a and 5b.

Although our system currently represents the state-of-the-art, the present version has the drawback of using stereo to select among the existing rectangles but of not using it to check for missed rectangles. Also there is a loss of accuracy in the determination of the disparities as a result of the robustness in the detection of the matched primitives. The rectangles are grouped features, and are thus primarily structural representations with low positional accuracy. The component lines of the rectangle only represent the structure among the underlying edges, not their exact positions. For obtaining accurate disparity, matching of more precisely located features, namely the edges, is required, using a system like the one described in [10].

Many steps can be taken however to improve the results. One way is to use more sensitive edge detectors on magnified portions of the image in small windows around the lines for precise detection and location of edges. We can also consider weaker edges near the noise level of the image, since we have an idea of the direction of the edges, their geometry (straight lines) and an approximate idea of their location.

6 Shape Description and Object Extraction

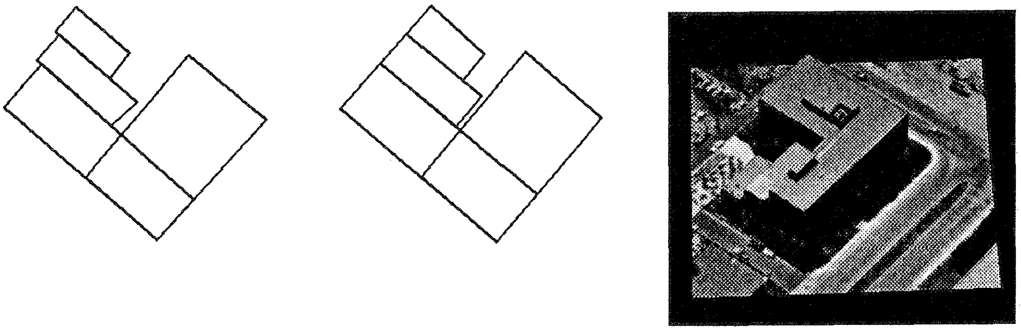
The detected groups usually correspond to parts of objects in the scene. We have to combine the grouped features into structures corresponding to the objects. The combination process automatically generates a shape description of the object in terms of the primitive shapes of the combined groups. In the case of simple buildings [30], the description of the shape is given by the lines (segments) along the outline of the objects. Some of these may be hypothesized to give partial descriptions of objects. In addition, the estimated height is given, computed from the width of the shadow. For runways [27] we give a description of the position, extent and orientation of the landing surface, together with a description of the evidence of markings.

In a more general manner our system for complex buildings identifies strong rectangle groupings which meet the height requirements of buildings in the scene. However, each rectangle grouping may not correspond to a separate roof since a roof shape could be a combination of rectangles. To extract an individual roof in the scene (object extraction), we consider possible combinations of the rectangles into structures which correspond to roofs. As the shape of a roof is described as a combination of rectangles, this process, in addition to *segmentation* also provides *shape description.*

The combination of the rectangle groupings is guided by reasoning based on the available 2-D and 3-D information. The visual reasoning carried out currently is primarily monocular, augmented by stereo as needed. The 2-D information is the geometrical relationships between the rectangles and the actual edges in the rectangle groupings. The 3-D information is obtained by stereo. The combination process is rule-based; the set of rules governing the combination of the rectangles (and the resultant structures) is defined on the 2-D and 3-D relationships among the rectangles.

In contrast to previous uses of monocular analysis, we work with more organized structures than lines and junctions. Also T-junctions, which are a key element in monocular analysis, can not be utilized for this application domain because of the presence of false T-junctions due to accidental alignments. The organized nature of the primitives used for processing brings more information to the monocular analysis than is available with just edge and junction information. The structural relationships considered in Mohan's work [44] are those of *subsumption* or *inclusion*, *merger-compatibility*, *occlusion* and *incompatibility.*

The geometrical relationships among the rectangles and their combinations form a graph which is a structural description of the objects in the scene in terms of the primitive rectangles. Struc-



(a) Left view (b) Right view (c) 3D model of building

Figure 5: Rectangles selected by stereo and model of building

tures in the graph which are not marked as subsumed, merged or incompatible are selected as the top level descriptions of the objects or object parts visible in the scene (the roofs in our example). The final structures are assigned heights from the disparity information obtained by stereo. The buildings are modeled by drawing walls straight down from the sides of the roofs to the plane below, be it another roof or the ground. The resulting model is displayed in figure 5c.

7 Integration of Information from Multiple Viewpoints

In the scenes analyzed by the systems mentioned above, many of the objects have restricted shapes and often the viewpoint is restricted. For some applications, it is necessary to integrate information extracted from images of a scene acquired from various viewpoints or acquired through various types of sensors. At the present time we are not aware of complete systems that provide information integration for photointerpretation tasks, and the existing techniques would have to be reviewed to determine the feasibility of relaxing viewpoint restrictions.

In our systems for instance, we detect two types of corners between the lines, L- and T-junctions. We currently do not investigate orthogonal trihedral vertices (OTVs) as few walls are visible, and those that are appear highly foreshortened and have shadows etc. near them making the OTVs difficult to detect accurately.

Also, T-junctions for traditional overhead urban aerial imagery do not have the usual interpretations of occlusion, as some of them would for oblique views. When the image plane is nearly parallel to the ground plane, the buildings may have wings and nearby objects like roads, etc. that are aligned to the building sides. In a top view, the sides of two different structures can create T-junctions in which the top line belongs to two different objects and is not occluding the stem. Therefore, the T-junctions are used to break the line belonging to the top of the T into sections.

If we continue to assume that we restrict the shape of the objects to rectangles, the most significant change is that right angles in the real world no longer necessarily project onto right angles in the image. The following changes to our systems, as suggested in [44], would have to be incorporated:

- L and T junctions will no longer be considered only for lines meeting at right angles. The lines may meet at any angle.
- Orthogonal Trihedral Vertices or OTVs will now be detected as for many views both the roofs and the walls will be visible.
- U-contours will be replace by *skewed* U-contours. The angle between the base and the parallel sides will no longer be restricted to right angles. However, since parallel lines still map to parallels, the sum of the angles between the base and the sides of a skewed-U will be constrained to lie near 180°.
- Rectangles will be replaced by parallelograms. Note that the parallelogram is composed of skewed-Us, parallels and lines in a fashion essentially the same as for rectangles.
- One possible new grouping to be considered is that of a "hinge." A hinge would correspond to two parallelograms which share a side. This creates two OTVs at the ends of the shared line. Given one parallelogram that is the projection of a side of a solid rectangular object, there is high probability that another side of the object, sharing a boundary, is visible. However, there could be cases where no hinge is present with a parallelogram due to occlusion. Therefore, a hinge would be a useful but not necessary grouping for detecting buildings.

The detection process for the grouped features, would essentially be the same as in our current system. However, due to the removal of some structural regularity (namely, right angles) the use of some geometric constraints may not apply.

The constraint satisfaction network will be similar to that in our current system with skewed-Us replacing U-Contours and parallelograms replacing rectangles. However, some new groupings like OTVs and Hinges will have to be incorporated. One possible source of complications is the presence of structures such as windows, textures, and ledges along walls (the roofs were relatively smooth). To add to this complication, some walls are reflective and reflect nearby objects.

8 Current Grouping Work

In the absence of computational theories of perceptual grouping we expect to see many approaches and techniques developed for specific applications. Invariably we can get bogged down in considering everything possible at all scales, and build complex and massive data structures. However, this is often unreasonable for mapping and photointerpretation tasks where the image content and typical resolutions quickly make some approaches unfeasible. Part of our current works thus, tries to contribute to the development of the computational aspects of perceptual organization. Specifically, we are working on developing a taxonomy of grouping operations and on the representation and computation of saliency.

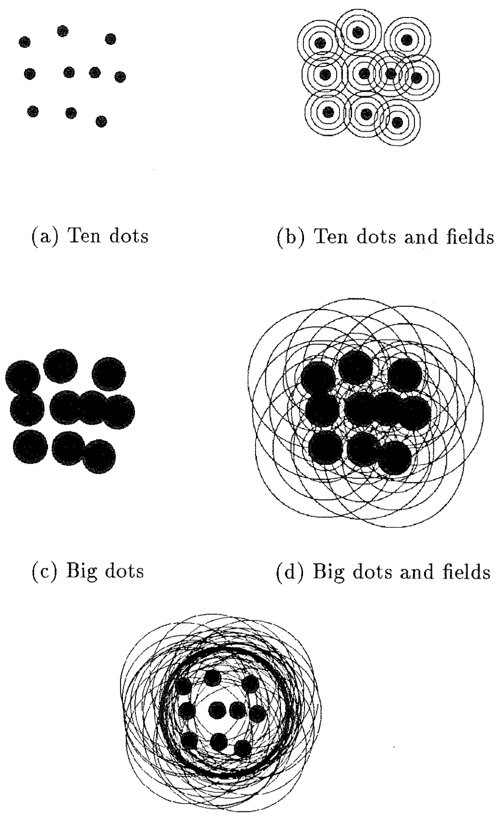
8.1 Grouping Fields

When a person states preference for one grouping over another, we believe they are expressing greater sensitivity to the saliency of a given attribute (or set of attributes). What makes a feature, a relationship, an attribute, salient?. What makes one feature more salient than another? The answer is that it can be grouped with another, or others, with the influence of each contributing to the strength of the percept. There are several related problems that must be solved to help understand the construction of a percept and its strength. We have discussed above our approach to perceptual grouping giving emphasis to structure and significance. The following notions continue this approach.

Consider the dots of radius r in figure 6a. If we are to decide whether to group them on proximity, we would consider how far they appear to be from one another. Suppose we are willing to allow a distance of up to three times their radii: It appears that these points are too far apart as their area of influence or grouping "attraction" in figure 6b indicates. Consider now the points at the same locations but having a radius of $3r$, shown in figure 6c: The points now appear closer and more likely to belong to a proximity group as their overlapping "influence", shown in figure 6d suggest. We call the of influence of a visual feature its *grouping field* with respect to a grouping operation, in this case *proximity*. The shape, extent and strength of the fields are defined as a function of the properties of the visual features (or objects), such as size, shape, color, and geometric complexity, much like the gravity field in Nature is a function of mass and distance. Note that we would judge then the distance among cars in a parking lot by the distances among the location of the cars' center of mass, and the extent of their proximity grouping fields to be a function of their size. Note also that a string of cars would not result in a single group but in many alternate colinear groups. If we restrict one car per group we would obtain a string of groups.

The most elementary proximity grouping operation, called Px0D, is then defined as a function of the distance between the center of mass of the elements to be grouped. In figure 6e the group results from combining overlapping fields. Each newly admitted element, shifts the center of mass of the group to a new location. Note that the resulting group can in turn be represented by a dot (located at the center of mass of the group). In this example the fields have the elementary interpretation of representing whether or not a feature is being influenced or not (within the field of) by a nearby feature. As we incorporate more feature (or group) attributes, the degrees of freedom or dimensionality of the operations increase, and are denoted Px1D, Px2D, and so on.

The most elementary collinear grouping operation, called Co0D, takes into account only the orientation of the features. In this example we apply Co1D to the lines shown in figure 7a (orientation and lateral tolerance to allow for distortions and inaccuracies). Lines falling within the area of influence of another line are grouped together. The collinearity groups are then represented by abstract ribbons having length, width and orientation derived from weighted contributions of the elements in the group (see figure 7b). The resulting group can be represented by a line (as in figure 7 and participate in other Co0D grouping operation at a higher level of the hierarchy, or we can use its orientation and width (the lateral tolerance used) in a Co2D grouping operation. This is equivalent to the grouping process that helped detect the runways [26] shown earlier in figures 3 and 4.



(a) Ten dots

(b) Ten dots and fields

(c) Big dots

(d) Big dots and fields

(e) Proximity influence

Figure 6: Proximity grouping

8.2 An Example

Consider analyzing a harbor or port complex. We wish to describe the buildings in the port facility, the transportation network around the facilities, and of course the pier areas and the ships in the area. We already discussed detection of buildings and transportation networks. What do we need to know about port and harbor facilities to detect the piers and describe the ships? That the planning and design of port and harbor facilities is strongly dependent on the characteristics of the ships to be served, and the type of cargo to be handled [62]. To eventually describe the scene completely we would have to know a lot of things about the ships: Main dimensions (length, beam, draft), cargo-carrying capacity, cargo-handling gear, types of cargo units, shape, hull strength and motion characteristics, mooring equipment, maneuverability, and so on. To detect only the pier areas (where later we would look for ships) we only need the upper bounds on ship dimensions and the approximate image resolution. These parameters are easily available a priori and chiefly determine the extent and strength of the grouping fields associated with the features that we will use.

Figure 8a shows an image of a portion of the U.S. Navy facilities in San Diego. We know that we should expect to see mostly military ships that may require long term docking, thus allowing for double or triple docking. We know the image resolution and the approximate ship dimensions, thus we know the minimum size of the piers. The following steps are applied:

Locate Boundary between Land and Water: We detect the boundary between land and water regions automatically using our implementation of [50]. In this example we arbitrarily selected the largest region to represent the water region. Next we approximate these boundary by piecewise linear segments, shown in figure 8b using LINEAR, our implementation of [48].

Locate "land" Structures in Water: Contrary to many natural structures on shore, cultural structures appear highly geometric. We expect that most piers will appear as linear structures attached to the shore, and in the water. Their linearity indicates that the piers or portions of piers should be characterized by apars (parallel groupings). Ships are typically docked parallel and adjacent to the piers. We then expect that most of the line segments corresponding to sides of piers, sides of ships, shadows, and so on in the neighborhood of the piers would result in many local parallel groupings (apars). The limit on parallel groupings is a function of image resolution and ship dimensions. The apars in our example are shown as thin lines in figure 8b.

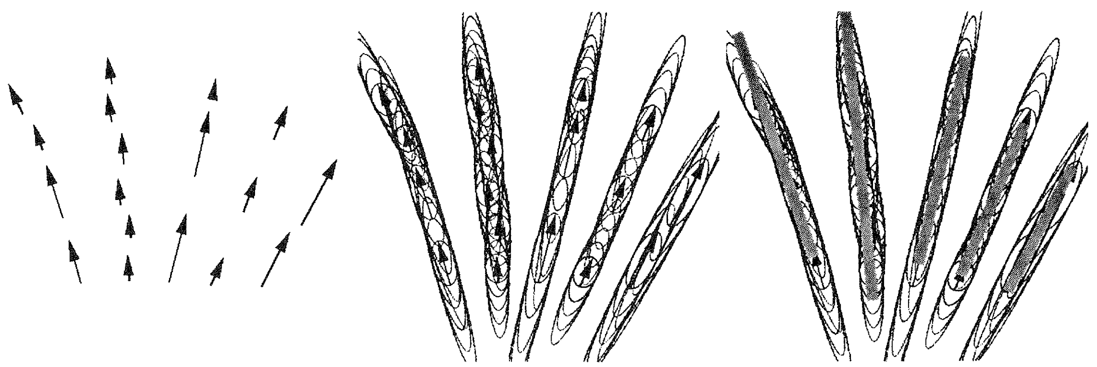
Detect Pier Areas: The apars are easily classified into "land" or "water" apar with respect to the water region. Subsequent processing operates on the land apars only. Next, we apply Px0D grouping to the land apars. The extent of the fields is task-dependent however it need to be only approximate. At the resolution in our example, the radii of the fields are roughly equivalent to the pier width plus the width of three destroyers on both sides of the piers, or about 16 pixels. Each apar (its center of mass) contributes a field. Subsequent contributions shift the center of mass of the group. We then select the groups so that apar membership is exclusive by extracting the possible groups in order of decreasing mass (number of apars). The resulting groups represent potential pier fragments (groups in figure 8b and arrows in figure 8c.) The representation of the resulting groups is the same as that of apars.

Next we apply ColD to the pier area fragments. The longest piers are about three times the length of a destroyer thus we allow the extent of the elliptic fields (see figure 8b) to be up to three times the length of the apars. The width is equivalent to the apar width (or group radius, 16 pixels in this example).

The result of the grouping, shown in figure 8d, is then represented, again, by apars and denote potential pier areas.

8.3 Saliency-ehancing operators

The second effort deals with saliency-enhancing operators capable of highlighting features which are considered perceptually relevant. These are introduced in [20]. They are able to extract salient curves and junctions and generate a description ranking these features by their likelihood of coming from the original

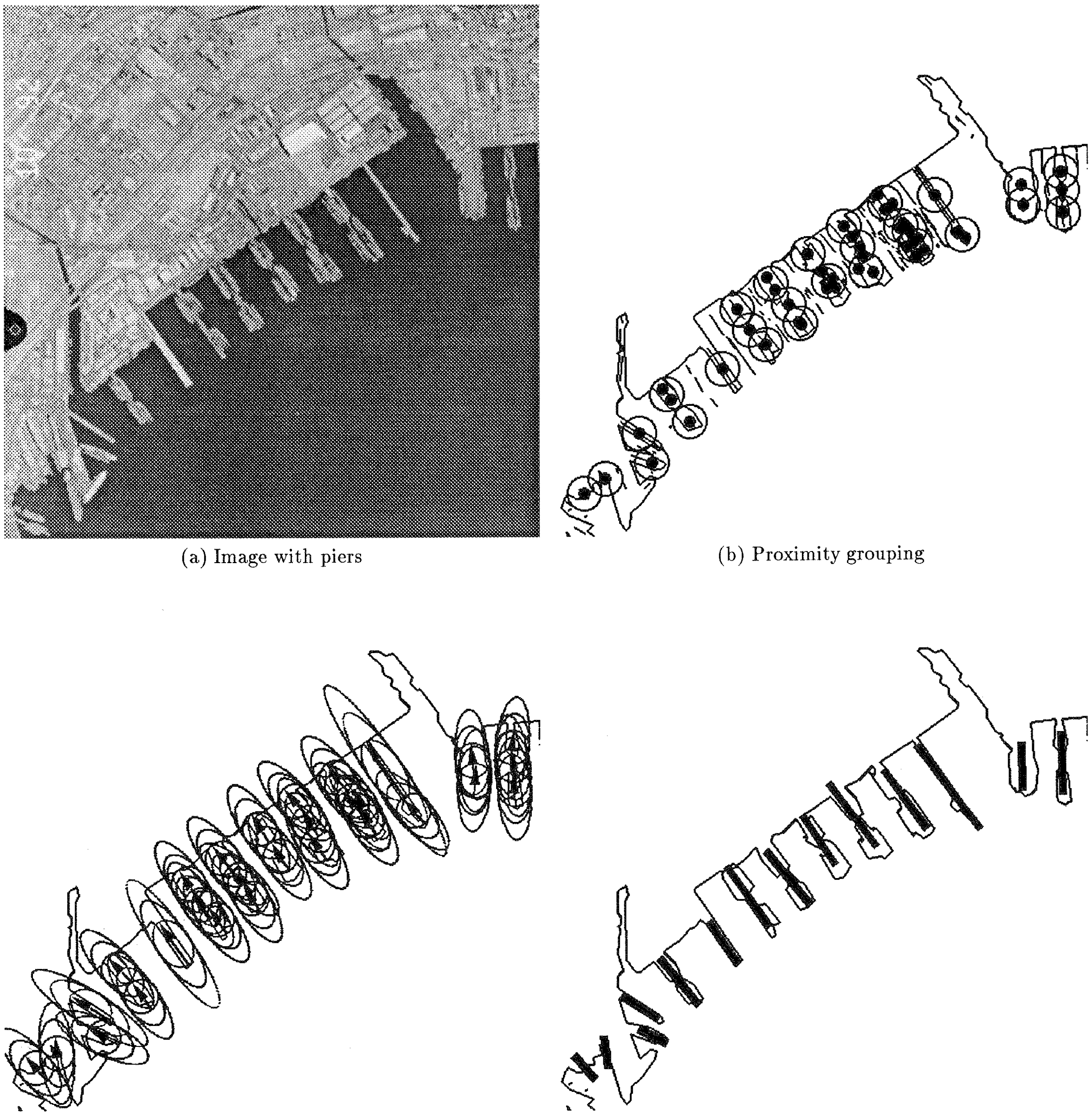


(a) Collinear lines

(b) Collinear influence

(c) Resulting groups

Figure 7: Colinearity grouping



(a) Image with piers

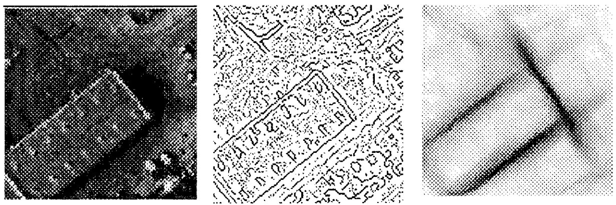
(b) Proximity grouping

(c) Collinearity grouping

(d) Pier areas

Figure 8: Low Resolution Harbor Scene

scene. They suggest the *global extension field* as means of describing the behavior of a curve segment, in terms of its continuation. They also show that a directional convolution of an edge image with the above field can produce useful descriptions. In this technique all operations are parameter-free, non-iterative and the processing is linear in the number of edges in the input image. As an example of this technique consider the small image containing a building in figure 9a. The intensity edges extracted from the image are shown in figure 9b. The saliency map constructed from these edges if shown in figure 9c.



(a) Image of building

(b) Intensity edges

(c) Saliency map

Figure 9: Computation of saliency

9 Bibliography

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

[2] D.H. Ballard. Form perception using transformation networks: Polyhedra. Technical Report TR 148, Department of Computer Science, University of Rochester, 1986.

[3] D.H. Ballard and C.M. Brown. *Computer Vision*. Prentice Hall, Inc., Englewood Cliff, New Jersey, 1982.

[4] D.H. Ballard, G.E. Hinton, and T.J. Sejnowski. Parallel visual computation. *Nature*, 306:21–26, November 1983.

[5] D. Blostein and N. Ahuja. A multi-scale region detector. *Computer Vision, Graphics, and Image Processing*, 45(1):22–41, January 1989.

[6] R. A. Brooks. Symbolic reasoning among 3-D models and 2-D images. *Artificial Intelligence*, 17:285–349, 1981.

[7] R. A. Brooks. Model-based three-dimensional interpretations of two-dimensional images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):140–150, 1983.

[8] G.A. Carpenter, M.A. Cohen, S. Grossberg, T. Kohonen, E. Oja, G. Palm, J.J. Hopfield, and D.W. Tank. Technical comments: Computing with neural networks. *Science*, 235, March 1987.

[9] C.-K. R. Chung and R. Nevatia. Recovering LSHGCs and SHGCs from stereo. In *Proceedings of the DARPA Image Understanding Workshop*, pages 401–407, San Diego, CA, January 1992.

[10] S.D. Cochran. *Surface Descriptions using Stereo*. PhD thesis, University of Southern California, August 1990.

[11] S.E. Fahlman and G.E. Hinton. Connectionist architectures for artificial intelligence. *IEEE Computer*, pages 100–109, January 1987.

[12] S.E. Fahlman, G.E. Hinton, and T.J. Sejnowski. Massively parallel architectures for AI: NETL, Thistle, and Boltzman machines. In *Proceedings, National Conference on A.I.*, Menlo Park, C.A., 1983. American Association for A.I., William Kafman, Inc.

[13] J.A. Feldman and D.H. Ballard. Connectionist models and their properties. *Cognitive Science*, 6:205–254, 1982.

[14] M.A. Fischler and R.C. Bolles. Perceptual organization and curve partioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(1):100–105, January 1986.

[15] P. Fua and A. J. Hanson. Using generic geometric models for intelligent shape extraction. In *Proceedings of the DARPA Image Understanding Workshop*, Los Angeles, CA, Feburary 1987.

[16] W.R. Garner. *The Processing of Information and Structure*. Erlbaum, Potomac, M.D., 1974.

[17] W.R. Garner and D.E. Clement. Goodness of pattern and pattern uncertainity. *Journal of Verbal Learning and Verbal Behavior*, 2:446–452, 1963.

[18] S.L. Gazit and G. Medioni. Multi-scale contour matching in a motion sequence. In *Proceedings of the DARPA Image Understanding Workshop*, May 1989.

[19] S. Geman and D. Geman. Stochastic relaxtation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, November 1984.

[20] G. Guy and G. Medioni. Perceptual grouping using global saliency-enhancing operators. In *Proceedings of the 2nd International Conference on Pattern Recognition*, The Hague, Netherlands, 1992.

[21] A. Hanson and R.E. Riseman. *VISIONS: A Computer System for Interpreting Scenes*. Academic Press, New York, 1978.

[22] M. Herman and T. Kanade. Incremental reconstrucution of 3d scenes from multiple, complex images. *Artificial Intelligence*, 30:289–341, 1986.

[23] J.E. Hochberg and E. McAllister. A quantitative appraoch to figural "goodness". *Journal of Experimental Psychology*, 46:361–364, 1953.

[24] J.J. Hopfield and D.W. Tank. "Neural" computation of descisions in optimization problems. *Biological Cybernetics*, 52:141–152, 1985.

[25] A. Huertas. Using shadows in the interpretation of aerial images. Technical Report 104, USC-ISG, October 1983.

[26] A. Huertas, W. Cole, and R. Nevatia. Using generic knowledge in analysis of aerial scenes: A case study. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1642–1648, Detroit, MI, August 1989.

[27] A. Huertas, W. Cole, and R. Nevatia. Detecting runways in complex airport scenes. *Computer Vision, Graphics, and Image Processing*, 51(2):107–145, August 1990.

[28] A. Huertas, R. Mohan, and R. Nevatia. Detection of complex buildings in simple scenes. Technical Report IRIS#203, Institute for Robotics and Intelligent Systems, University of Southern California, September 1986.

[29] A. Huertas and R. Nevatia. Detection of buildings in aerial images using shape and shadows. In *Proceedings of IJCAI*, pages 1099–1103, August 1983. Karlsruhe, W. Germany.

[30] A. Huertas and R. Nevatia. Detecting buildings in aerial images. *Computer Vision, Graphics, and Image Processing*, 41(2):131–152, February 1988.

[31] D.J. Jacobs. GROPER: A grouping based recognition system for two dimensional objects. In *Proceedings of the IEEE Computer Society Workshop on Computer Vision*, Miami Beach, Florida, December 1987.

[32] R.E. Kelly, P.R.M. McConnell, and S.J. Mildenberger. The gestalt photomapping system. *Journal of Photogrammetric Engineering and Remote Sensing*, 1977.

[33] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[34] K.A. Lantz, C.M. Brown, and D.H. Ballard. Model-driven vsion using procedural description: Motivation and application to photointerpretation and medical diagnosis. In *Proceedings 22nd International Symposium, Society of Photo-optical Instrumentation Engineers*, San Diego, CA, August 1978.

[35] E.L.J. Leewburg. Quantification of certain visual pattern properties: Salience, transparency, similarity. In Leewburg and Buffart, editors, *Formal Theories of Visual Perception*. Wiley, Chichester, U.K., 1978.

[36] H.S. Lim and T.O. Binford. Stereo correspondence: A hierarchical approach. In *Proceedings of the DARPA Image Understanding Workshop*, pages 234–241, Los Angeles, February 1987. Morgan Kaufmann Publishers, Inc.

[37] D.G. Lowe. *Perceptual Organization and Visual Recognition*. Kulwer Academic Press, Hingham, MA, 1985.

[38] D.G. Lowe. The viewpoint consistency constraint. *International Journal of Computer Vision*, 1(1):57–72, 1987.

[39] D.G. Lowe and T.O Binford. Perceptual organization as a basis for visual recognition. In *Proceedings of AAAI-83*, Washington, D.C.,, August 1983.

[40] D. Marr. *Vision*. W.H. Freeman and Co., San Francisco, 1982.

[41] T. Matsuyama and V. Hwang. SIGMA: A framework for image understanding: Integration of bottom-up and top-down analyses. In *Proceeding of the Ninth International Joint Conference on Artificial Intelligence*, Los Angeles, California, August 1985.

[42] D.M. McKeown, W.A. Harvey, and J. McDermott. Rule-based interpretation of aerial imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(5):570–585, 1985.

[43] G. Medioni and R. Nevatia. Segment-based stereo matching. *Computer Vision, Graphics and Image Processing*, 31:2–18, 1985.

[44] R. Mohan. *Perceptual Organization for Computer Vision*. PhD thesis, University of Southern California, August 1989. Report IRIS 524.

[45] R. Mohan, G. Medioni, and R. Nevatia. Stereo error detection, correction, and evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(2), Feburary 1989.

[46] R. Mohan and R. Nevatia. Segmentation and Description Based on Perceptual Organization. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 333–341, San Diego, California, June 1989.

[47] R. Mohan and R. Nevatia. Using Perceptual Organization to Extract 3-D Structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(11):1121–1139, November 1989.

[48] R. Nevatia and K.R. Babu. Linear feature extraction and description. *Computer Vision, Graphics and Image Processing*, 13:257–269, 1980.

[49] R. Nevatia, K. Price, and G. Medioni. USC image understanding research: 1989–90. In *Proceedings of the DARPA Image Understanding Workshop*, Pittsburg, Pennsylvania, 1990. Morgan Kaufmann Publishers, Inc.

[50] R. Ohlander, K. Price, and R. Reddy. Picture segmentation by a recursive region splitting method. *Computer Graphics and Image Processing*, 8:313–333, 1978.

[51] Y. Ohta and T. Kanade. Stereo by intra and inter-scanline searching using dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, March 1983.

[52] S.E. Palmer. The psychology of perceptual organization: A transformational approach. In J. Beck, B. Hope, and A. Rosenfeld, editors, *Human and Machine Vision*, pages 269–339. Academic Press, New York, NY, 1983.

[53] L. Quan, R. Mohr, and E. Thirion. Generating the initial hypothesis using perspective invariants for a 2D image and 3D model matching. In *International Conference on Computer Vision*, Florida, December 1988.

[54] G. Reynolds and J.R. Beveridge. Searching for geometric structure in images of natural scenes. In *Proceedings of the DARPA Image Understanding Workshop*, Los Angeles, C.A., Feburary 1987.

[55] D.E. Rumelhart, McClelland, and the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructures of Computing.* M.I.T. Press, 1986.

[56] H. Sato and T. Binford. BUILDER-I: a system for the extraction of SHGC objects in an edge image. In *Proceedings of the DARPA Image Understanding Workshop*, pages 779–791, San Diego, California, January 1992.

[57] H. Sato and T. Binford. On finding the ends of SHGCs in an edge image. In *Proceedings of the DARPA Image Understanding Workshop*, pages 379–388, San Diego, California, January 1992.

[58] F. Stein and G. Medioni. Toss - a system for efficient three dimensional object recognition. In *Proceedings of the DARPA Image Understanding Workshop*, Pittsburgh, Pennsylvania, September 1990.

[59] K.A. Stevens. Computation of locally parallel structure. *Biological Cybernetics*, 29:19–28, 1981.

[60] A. Triesman. Perceptual grouping and attention in visual search for features and objects. *Journal of Experimental Psychology: Human Perception and Performance*, 8(2):194–214, 1982.

[61] A.P. Witkin and J.M. Tenenbaum. On the role of structure in vision. In J. Beck, B. Hope, and A. Rosenfeld, editors, *Human and Machine Vision*, pages 481–543. Academic Press, New York, NY, 1983.

[62] P. Wright and N. Ashford. *Transportation Engineering: Planning and Design.* John Wiley and Sons, 1989.

[63] S.W. Zucker. Computational and psychophysical experiments in grouping: Early orientation selection. In J. Beck, B. Hope, and A. Rosenfeld, editors, *Human and Machine Vision*, pages 545–567. Academic Press, New York, NY, 1983.