Qualitative Building Detection from Monocular Images Based on Linquistic Context*

Zhongfei Zhang
zhongfei@cedar.buffalo.edu

Rohini Srihari
rohini@cedar.buffalo.edu

Center of Excellence for Document Analysis and Recognition (CEDAR)
State University of New York at Buffalo, Buffalo, NY 14228-2567

Abstract

This paper addresses an important and practical problem in computer vision — qualitative object detection/recognition. In particular, it discusses the problem of qualitative building detection in aerial images. The approach we propose in the paper, due to its independence of 3D site models and/or camera pose/calibration information, complements the model based approaches in the literature in the sense that it exploits the linquistic context in solving for the problem of qualitative building detection, and uses this contextual information extensively to guide the process of detecting buildings in the image domain. A knowledge supervised perceptual grouping algorithm based on the input linquistic context is presented, and is shown to be reasonably robust in experiments using RADIUS model board images.

1 Introduction

This paper addresses an important and practical problem in computer vision — qualitative object detection/recognition. In particular, it discusses the problem of qualitative building detection in aerial images.

Building detection in aerial images is an important yet difficult problem. Solutions to this problem find many application domains such as cartography, intelligence analysis, surveillance monitoring, and target recognition. Due to its practical interests, there have been several attempts to solve this problem [5, 6, 13, 12, 19]. Recently, with the progress of the RADIUS program [1], several more successful building detection systems have been developed. Some of these may be found in [11, 3, 4, 9, 14, 7]. All of these approaches assume that there are site models available for interpretation of the 2D images.

*This work was supported in part by ARPA Contract 93-F148900-000.

In recent research in image understanding and computer vision, context information has received more and more focused attention [18, 17, 16]. Unlike most existing building detection algorithms, which detect buildings based on site models, the algorithm proposed in this paper is different in the sense that it exploits the linquistic context in solving for the problem of qualitative building detection, and uses this contextual information extensively to guide the process of detecting buildings in the image domain; it is completely independent of any 3D site models and/or camera pose and calibration information. While model based building detection is certainly important, and has found many application scenarios (e.g. model supported exploitation [1]), there are circumstances where 3D site models may not be available. Moreover, people may only be interested in approximate segmentation of building areas in 2D images based on qualitative annotation, rather than precise delineation of the building boundaries, or reconstruction of the precise 3D building wireframe models. A typical scenario is as follows. An image analyst (IA) annotates a set of images which do not have corresponding 3D site models. The annotation consists of a brief text or speech regarding a region of interest in an image to identify targets (e.g. buildings), and to describe any available functionalities of those targets, such as what this building is used for, what facilities this building may have, etc. A typical description is shown in Fig. 1. A building detection system is then expected to locate where the mentioned buildings are in the image, roughly segment out those buildings in the image, and mark them with any annotated functionality information. The next time when this IA or other IAs retrieve the image, this information may be used to either update the annotation of this image, or serve as "site folders" (background information) for annotation of other related images. Since annotation through IAs' speech or text typically can only be expected to provide qualitative information (e.g. a description of

rectangular buildings as opposed to a precise, geometric description of the buildings, which may not be available at this point to the IAs), the building detection system described here constructs models (used to detect buildings) based purely on qualitative information.

Qualitative building detection has two implications. First, there is no precise and/or quantitative building models; rather, the model available is qualitative (we call it *generic* model). Second, building detection here refers to approximate segmentation of the building area in an image, disregarding any "details" of the building boundaries (e.g. a building may have some tiny protrusion or indention). Since these "details" normally cannot be substantiated in a qualitative description (e.g. a rectangular building with a tiny protrusion on one side is still called a rectangular building), the qualitative building detection proposed in this paper ignores all these "details".

The kernel of our qualitative building detection algorithm is based on knowledge supervised perceptual grouping. Perceptual grouping has been extensively used in grouping spatial data into semantically meaningful entities [10, 15]. A typical recent work is using Bayesian networks by Sarkar and Boyer [15]. In particular, the techniques of perceptual grouping have been widely used in building detection [13, 7]. Jaynes et al [7] used graph theory in perceptual grouping. Mohan and Nevatia [13] explored stereo images for perceptual grouping and used a constraint satisfaction network in the process of perceptual grouping. Unlike the previous work, here we use perceptual grouping in a single image, and conduct the grouping by constantly exploring context of tokens in the raw image.

This paper is organized as follows. The next section briefly introduces the qualitative input to this building detection system. Then a knowledge supervised perceptual grouping technique based on generic models is presented. Finally, the algorithm is followed by experiments and conclusions.

2 Knowledge Supervised Perceptual Grouping (KSPG)

Given a generic model and a qualitative viewing direction, qualitative building detection is based on knowledge supervised perceptual grouping (**KSPG**). There are three levels for this grouping. Given a *raw image*, a line-finding algorithm is applied to the raw image to obtain a *token image*, in which each token is a line segment composed of a set of connected edge pixels in the same direction. Due to the generic imperfection of line-finding algorithms, and the fact that there is always "noise" in an image such as shadows, the tokens need to be grouped to form a semantically correct segmentation. In order to give a correct building detection, perceptual grouping has to be applied to detect all the semantically correct lines (i.e. building image boundaries), and ignore all the "noisy" lines such as those cultural and textural lines inside the building boundaries, and the shadow lines. The result of this detection is called the *label image*. The difficulty of perceptual grouping relies on the fact that it is difficult to make a decision about whether or not to group two tokens together, or to discard a token when in grouping, if perceptual grouping is *only* performed at the token level. This decision can only be correctly made if appropriate context knowledge is provided. Our system exploits two sources of context knowledge.

```
This image depicts the Buffalo
Development and Manufacturing Complex.
In the western half of Area Delta
is the Baird Research Institute.
It consists of four buildings.
Label the leftmost long rectangular
building as Kelly Laboratory.
Label the large two-storied building
as the Headquarters building.
Label the L-shaped building as the
Operations Center.
Label the small square building as
the Communications Center.
```

Figure 1: Example speech as qualitative input to the building detector.

One is from the qualitative parameters extracted from the description of the linquistic input to the building detection system. This information contains qualitative description of building shapes which gives the generic model for perceptual grouping (e.g. rectangular, L-shaped, etc.), qualitative viewing directions (e.g. nadir view, oblique view, etc.; the default is nadir view in our system.), number of stories (e.g. two-story), and shadow information (e.g. yes/no shadow; the default is yes). Other qualitative information may also be used as input, such as qualitative description of the location (e.g. this building is located in the southwest corner), but it is not necessary. To give an example about how **KSPG** works, if the input parameters are rectan-

gular and nadir-view, **KSPG** searches for a rectangle; if the input parameters are rectangular and oblique-view, **KSPG** searches for a non-orthogonal parallelogram.

The other source of context knowledge comes directly from the raw image itself. The raw image provides rich contextual information, which may be used to guide perceptual grouping. For example, assume that there are two line tokens lying in approximately the same direction with a small gap between them. They may be two parts of the same physical line representing the same cultural boundary, or they may represent two separate cultural boundaries. Merging them together based only on proximity at the token level may lead to wrong decision and incorrect result. Consequently, the contextual information in the raw image should be constantly consulted in every step when tokens need to be merged or to be discarded in perceptual grouping. This process of going back to check the context of tokens in the raw image in perceptual grouping is called *reinvestigation* in this paper. We believe that reinvestigation is necessary in perceptual grouping.

2.1 Reinvestigation

Reinvestigation is the essential part of **KSPG**. There are two types of reinvestigation. One is called *longitudinal merge and extension.* Fig. 2 shows how it works. Fig. 2(a) shows two tokens lying in the same direction separated by a close gap. We assume that if they belong to the same physical line, the intensity values of the pixels at the interior sides of the two tokens and the gap should possess continuous and consistent statistical properties, as shown in Fig. 2(b); on the other hand, if they represent two different physical lines, there should be a discontinuous and inconsistent change of statistical properties of intensity values in the gap between the two tokens, as shown in Fig. 2(c). The statistical properties may be defined in various forms. In this implementation of the system, we use the median filtered intensity values along a line constantly aside the tokens. This figure shows the rationale of longitudinal merge. Similarly, the same idea applies to longitudinal extension where a token may be extended as far as the actual physical line this token represents ends.

The other type of reinvestigation is called *lateral support.* This applies to the scenario of completing a building hypothesis with one missing token. For example, three sides of a rectangular building rooftop have been found with the fourth side absent in the token image. Lateral support refers to the process of searching and verifying the existence of the token corresponding to this missing side in the raw image. Again, similar statistical analysis is used to find the missing token in order to complete the building hypothesis.

2.2 Preprocessing

Before applying **KSPG**, some preprocessing is necessary in order to obtain an initial token set and token image, and to roughly locate where the building is in the image. There are three major steps of preprocessing:

- **Token Extraction:** A line finder is applied to the raw image to obtain an initial extraction of line tokens in this image. The result is an abstract token set, and a token image which shows all the detected tokens in an image plane. We use UMass Fast Line Finder [8] and have ported this line finder to our system with a new graphical interface developed by ourselves. Fig. 4 shows part of RADIUS M7 image with line tokens superimposed on it by application of this line finder.

- **Histogram Analysis and Dominant Token Set Determination:** A histogram analysis is conducted based on normalized token lengths of the whole token set w.r.t. the orientation of each token. Each token is normalized in terms of its length by the longest token of the whole token set, and is put into a bucket corresponding to its orientation. Fig. 6 shows this histogram of the token set in Fig. 4. After this histogram is obtained, all the tokens corresponding to the global peak of the histogram are taken out to form a token subset, called a *dominant token set*, and each token in the dominant token set is called a *dominant token.*

- **Reference Point Set:** As explained in the next subsection, the hypothesis generation and verification of **KSPG** assumes a known interior point of a building candidate. Thus, as the last step of the preprocessing, a reference point set is generated such that some of the points in this set are located inside the building boundaries that need to be detected. In order to generate this reference point set, each line token is represented as a mass point with the normalized length as its mass and the midpoint as its location. Then a nearest-neighbor based clustering algorithm [2] is used to obtain a final distribution of the mass cluster centers in the image.

For those images containing man-made buildings, building scenes are often filled with long line tokens, and therefore the clustered mass centers normally fall inside the building boundaries.

2.3 Hypothesis Generation and Verification

Given a reference point, the process of hypothesis generation and verification of **KSPG** is conducted through the following three major steps:

- **Distance Map Construction:** The perpendicular distance from a reference point to each dominant token is calculated and associated with this token. Then all dominant tokens at the both sides of the reference point are sorted in terms of their distance to this reference point. The spatial layout of the sorted dominant tokens and the reference point is called a *distance map* as illustrated in Fig. 7.

- **Longitudinal Merge and Extension:** Tokens with similar distances to the reference point may be merged together to form a longer and more complete line token if they represent the same physical boundary line, or may be discarded if they do not represent the same building boundary, based on reinvestigation analysis (see Section 2.1). Similarly, a single line token may be extended in both directions as far as it can as long as there is contextual support, as explained in Section 2.1. The distance map is dynamically updated in the process of longitudinal merge and extension. This step completes detection of all the potential dominant boundary tokens of a building delineation.

- **Dynamic Matching:** The distance map is dynamically searched to match against the generic model. The dynamic matching process is a repetition of hypothesis generation and verification. A *candidate token set* is defined as those dominant tokens that may match the generic model. For example, a candidate token set of a rectangular model is any two dominant tokens at the both side of the reference point with similar lengthes. A hypothesis is *generated* when a candidate token set is found. The hypothesis is *verified* if there is contextual lateral support (see Section 2.1). A candidate is said to be *valid* if the hypothesis is verified. Note that there may be more than one valid candidates when the whole distance map is matched against the generic model. Some of them may be semantically true candidates, e.g. a two-storied building (such as one of the buildings in Fig. 5) with the reference point falling inside the second story building boundaries should return two candidates both being semantically correct; some of them may be false candidates, e.g. a rectangular building may have a rectangular chimney on the rooftop, and the reference point happens to be inside the chimney boundaries; then both the chimney boundaries and the building boundaries would be returned as valid candidates, but the chimney boundaries are clearly not the true candidates. In general, a metric similarity function is used to define the degree of matching between a valid candidate and the generic model, and only those valid candidates with similarity function values above a given threshold are considered as true candidates. For instance, parallelism is used to measure the degree of matching for rectangular generic model. By default, the true candidate with the largest similarity function value is returned as the final detection result. In the case of multiple storied buildings, however, candidates' size information is further considered and the outermost one is returned as the final detection result. This is another example to exemplify how qualitative contextual knowledge (here the input of multi-storied buildings) may be used to supervise the perceptual grouping.

Note that **KSPG** detects buildings based on segmentation in terms of generic models. If an image is from nadir view or approximate nadir view, **KSPG** returns the building boundaries; if the image is from oblique view, **KSPG** may only detect the roofs of the buildings, which is still considered as correct detection, because the algorithm does not account for the "height" dimension.

3 Experiments, Limitations, and Evaluations

The whole system is implemented under the interface shown in Fig. 3.

KSPG based qualitative building detection has been extensively tested on RADIUS model board images. Fig. 5 shows an example of the performance of the system. Given the qualitative linquistic input shown in Fig. 1, all the four buildings are detected as shown in Fig. 5. Note that the long rectangular building on the left and the two-storied rectangular

building on the right both have a little protrusion in the right sides of their building boundaries. These protrusions were not delineated as building boundaries because the qualitative building detection does not account for these details.

Fig. 8 is the detection result of our system for all the rectangular buildings in another RADIUS model board image (M34). This time the qualitative input is only a rectangular model. We do not even specify how many rectangular buildings are in the image. Note that there are actually seven rectangular buildings, and more importantly, those seven rectangular buildings have gable roofs which do not fit the generic rectangular model of a flat roof that we have implemented. However, since the view angle is not very oblique, the gable roofs are not very obvious. We intentionally apply our building detection algorithm to this image to test the robustness of the algorithm. Our system detected five out of the seven rectangular buildings. Fig. 9 shows the initial line token extraction of the area with four of the seven buildings. It can be seen that the building boundary tokens of the two buildings that failed to be detected do not satisfy the rectangular generic model well because the lateral tokens were not recognized as being parallel. On the other hand, this experimental result shows that our qualitative building detection algorithms works reasonably robust in the cases where actual buildings do not exactly fit the generic models. Also note that the three parking areas were not detected as buildings because our system assumes presence of shadow context as a default parameter.

Due to the "qualitative" nature of **KSPG** defined in the context of this paper, as well as the different assumptions of **KSPG** as compared with those of other algorithms, it is difficult to have an objective performance comparison of **KSPG** with others. Therefore, we use "self-evaluation" to show the performance of **KSPG**. The "self-evaluation" is characterized by the two statistics of *detection rate* and *false positives*. The detection rate is defined as the ratio of the total nubmer of *detected* buildings and the total number of the buildings that are referred by the qualitative linquistic input over all the images tested. The false positives are defined as the ratio of the total number of buildings that are *incorrectly detected* (i.e., the total number of the reported buildings detected by the algorithm but actually not real buildings) and the total number of buildings that are referred by the qualitative linquistic input over all the images tested.

Five RADIUS model board images are tested for **KSPG** for certain building types (generic models). The detection rate is 76.5%, and the false positives are 2.9%. The low false positives are owing to the reinvestigation technique employed in **KSPG**. There are two main reasons that contribute to the scenarios in which buildings fail to be detected.

- *Low image quality:* When an image quality is low, many important line tokens may be missed, and much noise may be present in the token image returned by an line finder. Although the reinvestigation technique has the capability to "pick up" those line tokens that were missed by the line finder, this capability is limited under certain range of thresholds. Moreover, the technique of reinvestigation is based on the assumption that (at least part of) dominant tokens are present in the image. If a dominant token is completely missed, which is possible in a low quality image, the reinvestigation technique does not work, and the corresponding building may fail to be detected.

- *Limitation of generic models:* Generic models that have been used in this system have limitations in terms of their descriptive nature. Many buildings may not exactly fall into any generic models. Some buildings are even difficult to be described by a generic model. Fig. 10 shows an example. This is a rectangular building. However, it has two protrusions on the roof, and even worse, one of the protrusions shares a wall with the main body of the building. This sharing of the wall prevents a whole side of the buiding boundary from having a complete dominant line token. Since the protrusion shares the same wall with the main body, reinvestigation is unable to "complete" the building boundary based on the rectangle generic model (because there is no edge information at all with the sharing part along the physical boundary). On the other hand, since this protrusion is significantly large, it cannot be ignored as a "detail" in the qualitative building boundary delineation. Hence, **KSPG** fails to detect this building. This is the main reason why the detection rate of **KSPG** is lower than those of some reported systems in the literature. Note that the difference here is that there is no quantitative models (e.g. 3D models) incorporated, and thus, the input knowledge to the perceptual grouping has much less assumed information in **KSPG** than those reported methods.

The first limitation may be overcome by further relaxing certain assumptions imposed to our current reinvestigation system, so that the capabilities

of reinvestigation may be further enhanced, e.g. to remove the assumption that all the dominant tokens need to be present. While this is doable and will certainly improve the capability and performance of this building detection system, the trade-off, however, is that the system would become more complicated to implement. The second limitation is more intrinsic due to the nature of this approach. As indicated in the beginning of this paper, the main purpose of this research is to show the capability and the potential to exploit the linquistic contextual information in vision applications. It should be noted that although this approach is limited to certain degree, in many applications, it is necessary to apply this technique in building detection.

Overally speaking, like many other perceptual grouping algorithms, the bottleneck of this algorithm still relies on many thresholds, starting from the line finder, to reinvestigation. However, **KSPG** shows an alternative avenue towards building detection in the scenario where no quantitative model information is available. Experimental results show that with the reasonable image quality, and with simple building types (so the generic models may be applied), **KSPG** performs reasonably well.

4 Conclusion

This paper has addressed an important, challenging, yet very realistic problem in computer vision — qualitative target detection — with the application to aerial image annotation. The application scenarios we have proposed here is quite different from those where most of the existing building detection systems in the literature can apply, which need quantitative models. We have investigated the problem of qualitative building detection, and have proposed an algorithm for approximately detecting building boundaries in a 2D image based on qualitative linquistic context. The main contribution of this work is to extensively exploit the contextual information through the linquistic input and implement it in knowledge supervised perceptual grouping (**KSPG**). We have defined the two types of context knowledge used in our system to guide this particular application of perceptual grouping, and have discussed the details of how the **KSPG** works. Experimental results show that this proposed algorithm works reasonably robust in RADIUS model board images. Our current direction of research includes to implement more sophisticated generic building models, and to test our system with more complicated building images.

References

[1] DARPA. *Radius Testbed System operations concept.* DARPA, 1993.

[2] R. Duda and P. Hart. *Pattern Classification and Scene Analysis.* John Wiley and Sons, 1973.

[3] P. Fua. Cartographic applications of model-board optimization. In *IUW*, 1996.

[4] Y. Hsieh. Design and evaluation of a semi-automated site modeling system. In *IUW*, 1996.

[5] A. Huertas and R. Nevatia. Detecting buildings in aerial images. *Graphics and Image Processing*, 41(2), 1988.

[6] R. Irvin and D. McKeown. Methods for exploiting the relationship between buildings and their shadows in aerial imagery. *IEEE Trans. on Systems, Man and Cybernetics*, 19(6), 1989.

[7] C. Jaynes, F. Stolle, H. Schultz, R. Collins, A. Hanson, and E. Riseman. Three-dimensional grouping and information fusion for site modeling from aerial images. In *IUW*, 1996.

[8] P. Kahn, L. Kitchen, and E. Riseman. Real-time feature extraction: A fast line finder for vision-guided robot navigation. *Trans. PAMI*, 12(11):1098–1102, 1990.

[9] C. Lin and R. Nevatia. Buildings detection and description from monocular aerial images. In *IUW*, 1996.

[10] D.G. Lowe. *Perceptual Organization and Visual Recognition.* Academic Press, 1985.

[11] J.C. McGlone and J.A. Shufelt. Incorporating vanishing point geometry into a building extraction system. In *IUW*, 1993.

[12] D.M. McKeown. *Mapping and Spatial Modelling for Navigation*, chapter Toward Automatic Cartographic Feature Extraction. Springer-Verlag, 1990.

[13] R. Mohan and R. Nevatia. Using perceptual organization to extract 3-d structures. *IEEE Trans. PAMI*, 11, IEEE.

[14] S. Noronha and R. Nevatia. Detection and description of buildings from multiple aerial images. In *IUW*, 1996.

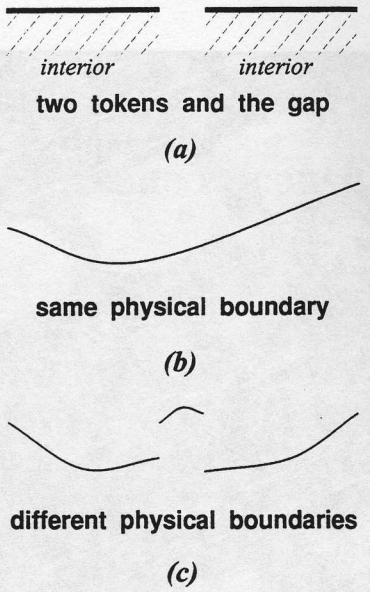

Figure 2: Longitudinal grouping. (a) two line tokens (b) if the two tokens represent the same physical boundary line, their interior parts should exhibit continuous and consistent statistical properties as those of the gap (c) if the two tokens represent different boundaries, their interior parts would have discontinuous and inconsistent statistical properties as compared with those of the gap.

[15] S. Sarkar and K.L. Boyer. Integration, inference, and management of spatial information using bayesian networks: Perceptual organization. *IEEE Trans. PAMI*, 15, 1993.

[16] R.K. Srihari, Z. Zhang, M. Venkatraman, and R. Chopra. Using speech input for image interpretation and annotation. In *Proceedings of Image Understanding Workshop*, 1996.

[17] Rohini K. Srihari. Use of Collateral Text in Understanding Photos. *Artificial Intelligence Review (special issue on integration of NLP and Vision)*, 8:409–430, 1995.

[18] Thomas M. Strat and Martin A. Fischler. Context-Based Vision: Recognizing Objects Using Information from Both 2-D and 3-D Imagery. *IEEE PAMI*, 13(10):1050–1065, 1991.

[19] V. Venkateswar and R. Chellappa. A framework for interpretation of aerial images. In *Proc. International Conference on Pattern Recognition*, 1990.

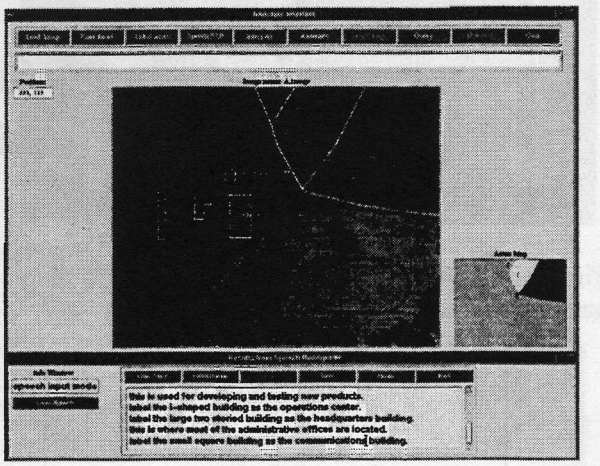

Figure 3: System interface. The upper window is the image interpretation window where the detected buildings, together with their annotations are overlayed onto the original image, and the lower window records the speech input.

Figure 4: An image of four buildings with initial line tokens superimposed on it.

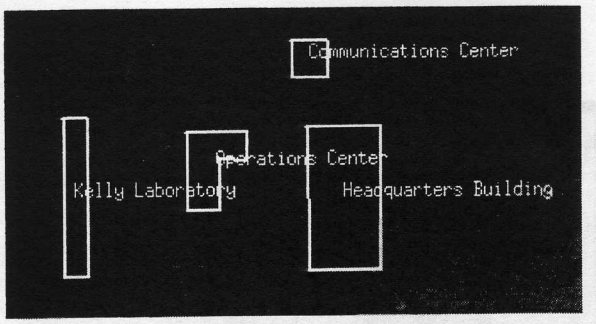

Figure 5: Qualitative building detection result of one RADIUS model board image.

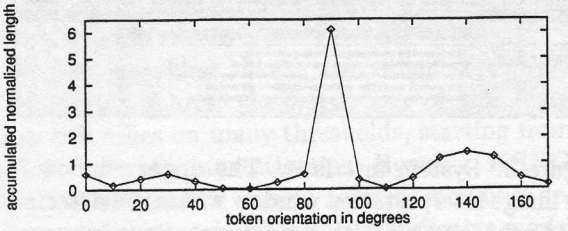

Figure 6: Histogram of the initial token set in Fig. 4.

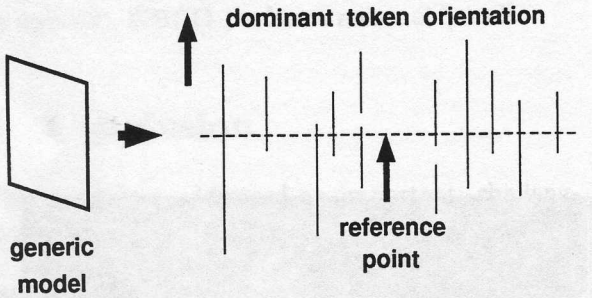

Figure 7: A generic model and a distance map in dynamic matching.

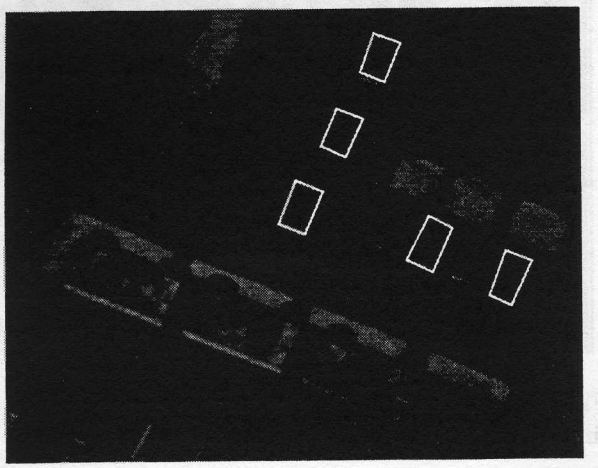

Figure 8: Detected rectangular buildings of image M34.

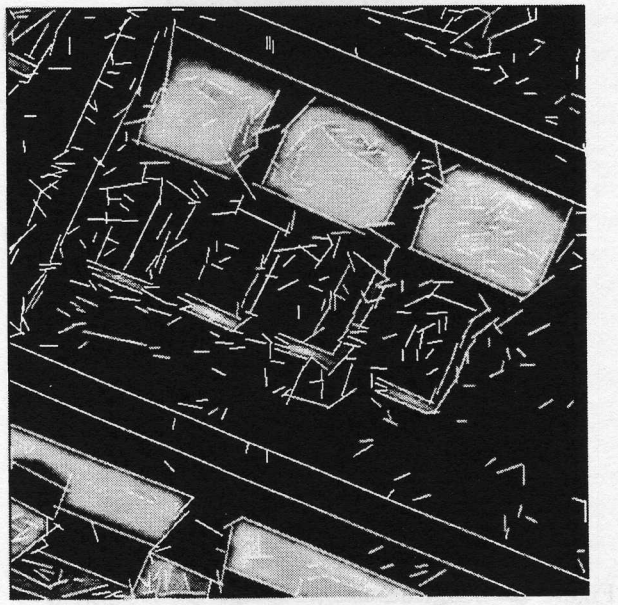

Figure 9: Part of M34 image with initial line tokens superimposed on it.

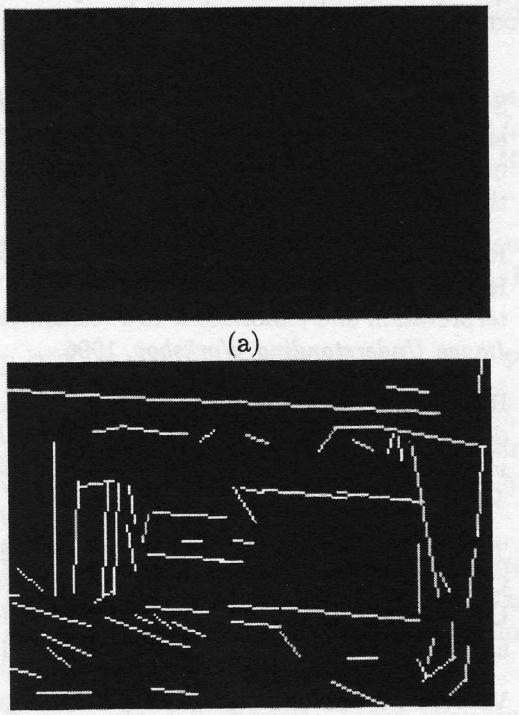

Figure 10: A building that is difficult to describe by a generic model. (a) the original image (b) the overlayed image with detected line tokens