MATCHING IN 2-D AND 3-D

R. Nevatia
Institute for Robotics and Intelligent Systems
University of Southern California
Los Angeles, Ca 90089-0273

KEY WORDS: Image Matching, Feature Matching, Scene Registration

PURPOSE:

This paper discusses issues and methods for matching in 2-D and 3-D. Several factors that affect the complexity of the task and the choice of the appropriate matching methodology . These factors include task characteristics such as whether the input is iconic or symbolic, 2-D or 3-D, the scene characteristics and the constraints from prior knowledge. Choices to be made consist of the representation level at which the matching is to be performed, whether it is in 2-D or 3-D, whether it is local or global, and the matching technique itself. The paper describes some general considerations which are then illustrated by two classes of specif problems, first being the problem of scene registration, the second being that of matching for depth estimation.

1. INTRODUCTION

Matching is of central importance for many image processing and understanding tasks. The process of object recognition essentially consists of matching stored object models with models derived from images. Process of change detection and map updating requires matching descriptions derived from new data with descriptions (maps or models) that have been constructed from analysis of earlier data. Extraction of 3-D information from a pair (or sequence) of images requires matching of corresponding (conjugate) points.

The complexity of the matching task and the appropriate strategy will depend on several factors listed below:

a) Iconic or Symbolic: Is the task to match entities in the image domain ({it iconic}), such as in stereo analysis, or to match an image with abstract models/ maps *symbolic* such as for change detection, navigation or object recognition? Certain kinds of methods, such as *area correlation* have no direct analog for symbolic matching. Note that even iconic matching may be performed by first extracting symbolic descriptions from images.

b) 2-D or 3-D input: The entities to be matched may represent 2-D or 3-D information, thus giving rise to four possible matching combinations. Images are usually 2-D though 3-D (or what some may call 2-1/2 D) is becoming increasingly available. Maps are often 2-D but more complex models of the scene as may be found in a GIS can be 3-D. When the objects to be matched are not of the same dimensionality, we need to compute a transformation between the two. Note that 3-D matching may be applicable even if we are trying to match 2-D images, as the underlying scene may be 3-D and and it may be necessary to make this 3-D structure explicit (as in stereo analysis).

c) Scene Characteristics: The observed scene content can vary from natural terrain to highly structured urban and suburban environments. Different matching techniques may be more appropriate for these different environments. In general, structured environments can be naturally described by abstract geometric shape whereas natural terrain may be better characterized by texture.

d) Constraints from prior knowledge: Complexity of the matching task can depend greatly on what constraints can be placed, say from the knowledge of camera geometry. For example, in stereo analysis, we can think of the task of computing the epipolar geometry or of utilizing given epipolar geometry for depth extraction.

In the following, we first examine some common issues related to matching under these varying conditions. Then, we

describe some illustrative systems for some specific tasks and discuss how these choice were made.

2. ISSUES IN MATCHING

Several important issues need to be considered in making the choice of a matching strategy for a particular task. These include the choice of representation level for matching, whether the matching is done in 2-D or 3-D, whether it is local or global and the method of matching itself. These are discussed below.

2.1 Representation Level:

Perhaps the most important consideration in matching is to determine the level (or levels) of representation at which matching is to be performed. Some possible levels are:

i) Direct pixel intensity (or color) values
ii) Point features such as local variance or edges
iii) Grouped features such as curves, sets of curves or regions
iv) High level features such as surfaces and volumes

Clearly, the features to be matched must be computable from the input data, must be invariant or quasi-invariant, and must be in the same form as the model or forms that can be derived from the model. The appropriateness of the level will depend on factors outlined in the introduction earlier, however, we can make some general observations about the choice:

i) Higher levels of representation require more complex algorithms: e.g. for intensity matching, simple correlation may be used but matching of surfaces may require graph matching procedures. Computational requirements at higher levels may be less due to the lower number of items to be matched; however, this may be compensated by the computational requirements of obtaining the higher level representations.
ii) Higher levels of representation are more distinct and thus the matching is likely to give less ambiguous results. However, more ambiguities may be present in the process of constructing the higher levels from the given image data in the first place.

2.2 2-D vs. 3-D Matching:

In general, matching in 3-D is more constrained and the 3-D features are more distinctive. Thus, if the input data is in a 3-D form, there are obvious advantages to performing the matching in 3-D. However, 3-D is not explicitly available in intensity images, in fact, extraction of 3-D may be an explicit goal of the matching process.

2.3 Local vs. Global Matching:

Another issue to consider is the extent over which matching procedure should be applied at a time. Local matching can be very precise, but ambiguous and the various local matches may not be consistent with each other. Global matching is more robust but not necessarily accurate in local areas unless a single transformation relates the sources to be matched. In general, we may need to make a compromise between the two extremes or proceed from local to global matches (or vice-versa) in stages.

2.4 Method of Matching:

Several techniques of matching are available, corresponding to the kinds of representations used. Three kinds of methods are listed below:

i) Area Correlation: here a single measure of match is computed by applying a transformation and matching the similarity between two sources.

ii) Feature Matching: This is a modification of the area correlation method. A certain transformation is applied to one source and the number (or amount) of matching features is computed. Determination of what constitutes a match is now more complex (i.e. when can two line segments be considered to match and to what degree).

iii) Structure Matching: Here a group of features is matched together, by considering not only individual feature properties but also some explicit relations between them (such as certain kinds of alignments, say parallelism). In general, these methods employ graph matching techniques.

A good survey of these issues and approaches can be found in ([6], [10]).

We now consider two specific kinds of tasks to illustrate the specific issues in matching and some kinds of techniques that have been used. First task is that of *scene registration* where one (or more) of newly acquired images need to be matched with a model (or map) of the site (constructed from earlier images or other sources); this task is important for purposes of change detection and model (map) updating. Second task is that of matching two or more images for the purpose of extracting 3-D models (maps). We will focus on cases where the input data consists of panchromatic intensity images and the scenes contain significant amount of man-made features rather than just natural terrain.

3. SCENE REGISTRATION

In this task, a new image needs to be registered with maps or models constructed from previous images. This operation is needed for several tasks such as detecting changes in the scene from the last time the models were constructed. Change detection is important for many civilian tasks such as map updating, urban monitoring and earth resource surveys and also for military tasks of observing significant infra-structure changes.

The complexity of this task varies greatly with the kind of

given input, the nature of the model and the constraints provided by the knowledge of the imaging parameters. We will consider three cases: matching range data with a digital terrain model (DTM), matching an image with a 3-D model with good camera orientation knowledge and a more general matching situation. All of these cases do share a common characteristic: the input image can be registered with the model by one global transformation (though the global transformation may be space variant to accommodate distortions of the sensor). The task of matching thus becomes that of estimating the parameters of the transformation. In general, we would need to estimate the interior and exterior camera parameters, though in most cases, some of the parameters may be known or constrained to be in a certain range.

3.1 Matching with a DTM

In this case, the model of the scene is simply an array of heights on a grid. The DTM itself may be constructed by stereo matching, by direct range sensing or by other means. Let us consider the case where the input image is also a range image (i.e. it contains height information).

We can consider both DTM and the range image to be like intensity images where the image value represents height rather than radiometric reflections. The search for transformation parameters can be reduced to search in the two-dimensional space of the ground plane. This search can be conducted by using conventional area cross-correlation methods. In such techniques, a measure of match is computed by some metric on point to point differences of height (or intensity) values: commonly used metrics are sum of the squares of differences or the cross-correlation coefficient. One array is translated relative to the other and the match metric computed for different displacements and the one with the best match is chosen. The search can be made more efficient by utilizing a pyramid of varying resolution images; coarse registration is achieved at the lower resolutions and the search at higher resolutions is confined to the range given by the lower resolution. Thus, high accuracy registration can be achieved efficiently. An analysis of the accuracy of this approach may be found in [7].

This technique is not directly applicable if the image is not a range image but a conventional intensity image as the height and intensity do not have a simple point to point correlation and the intensity image is also a function of additional camera parameters which need to be estimated. The author is not aware of systems performing such registration but believes that some form of feature matching as in the cases outlined below will be required.

3.2 Matching a 3-D model with known camera orientations

We now consider the task of registering an intensity image with a 3-D model of the scene (we will call it a *site model*. The site model itself may have been constructed from earlier images and other sources of information. The site model may contain various kinds of information, such as wireframe models of buildings in the scene, transportation networks, terrain heights and surface properties, depending on the application. The site model is, in general, a symbolic data structure and no point to point correspondence between it and an image is possible. Instead, we seek to find the transformation that projects the objects in the site model to corresponding objects in the image.

In many photogrammetric and remote sensing applications, the camera parameters are known with good precision. Internal camera parameters are known by a calibration procedure and external parameters are known from measurements on the sensor platform. Let us consider the case where camera parameters are known well enough so that the projection of the site model overlays the image to be registered well except for a translation in the image plane (the precision of the location of the platform may be lower than that of orientation). The task is now to find the correct translation as in section 3.1 above.

In this task, however, we still can not apply the method of pixel to pixel correlation as the projected model is not image like- it may only contain outlines, many parts of the scene may not be modelled at all and the projected structure does not have intensity values associated with it[1]. Instead, we need to compute some representations from both the image and the model that are similar and can be matched. The matching problem would be much easier if we could compute descriptions from the image at the high levels of abstraction that may be expected in the site model such as descriptions of buildings and transportation networks. However, such descriptions are difficult to infer reliably, so lower level features need to be considered.

We have developed a system for matching a site model to images where the dominant structures in the site model are polyhedral buildings [2]. In this case, linear line segments extracted from the image can provide sufficient features to match with line segments from projections of the models. Note that not all extracted lines will correspond to object boundaries and not all object boundaries will be so detected, but enough should be so that an overall match is possible. Figure 1 shows an example image which is to be registered with the model shown in Figure 2 (the figure shows the projection of the model from the expected view point). Figure 3 shows line segments extracted from the image of Figure 1. The model lines and the image lines can now be matched by selecting candidates from each set that collectivelly vote for the best match. Note, however, that the lines can not be matched on a point to point basis, as even small errors will cause the lines to not align precisely. Instead, we consider two line segments to match if they are within a certain distance of each other. Further, the contribution of

1. we could consider constructing an image from the model, but faithful reconstruction for the new imaging conditions is a difficult task, requiring detailed knowledge of the reflectance properties of the elements in the scene and of the imaging conditions

each line segment match is a function of the overlapped distance between the two. The best match is determined from the accumulated constributions. Figure 4 shows the accumulator array, with the peak indicating the best match. Figure 5 shows the result of registering the model to the image (the model boundaries are overlaid on the image). Details of the approach may be found in [14].

Note that the described processing only provides a transformation that relates the models to the image. This is not the same as actually matching building structures in the model with the buildings in the image. This step requires much more detailed processing. We need to examine how many of the model features can actually be found in the image and whether they are sufficient to confidently predict the presence of the building. Details of such processing are also given in [2].

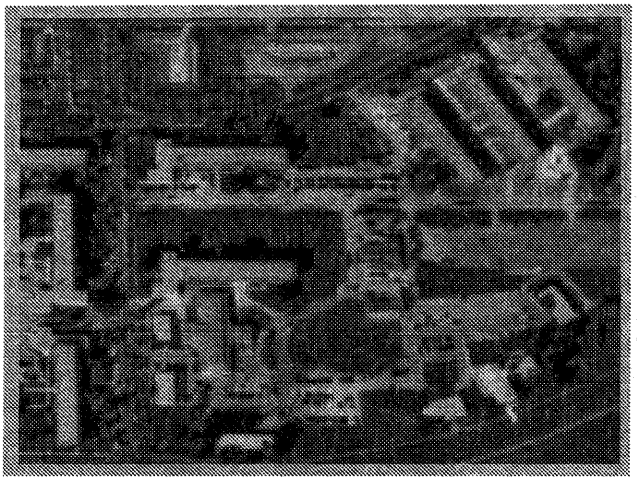

Figure 1 Image from Fort Hood, Texas

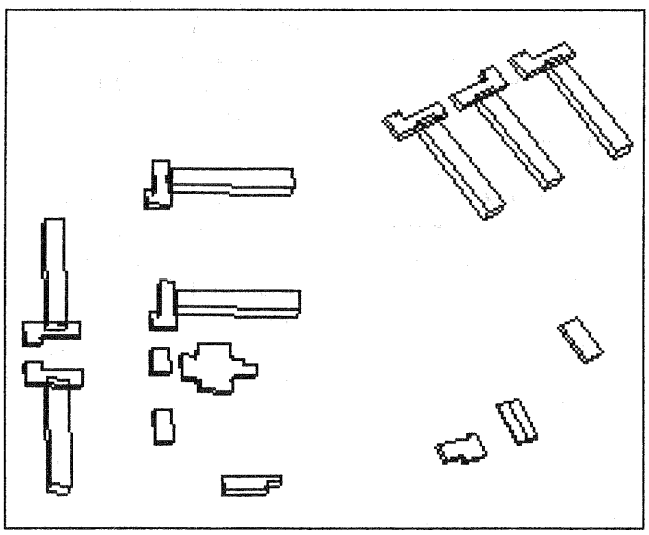

Figure 2 Model projection from expected viewpoint

3.3 Unknown Camera Orientations

If several parameters of the camera orientation are not known, the process of finding the best transformation as described above becomes much more complex. Instead of searching for two unknowns, we may need to search for five or more unknowns. In principle, the search can still be conducted as above, by hypothesizing different transformation parameters and computing a match score, but search in a five dimensional space may become prohibitive. An alternative is to use *alignment* [11] techniques. Here, a transformation (alignment) is computed from a small number of feature matches; the transformation can then be used to verify matching of the remaining features. The minimum number of features needed to estimate the transformation depends on the nature of features (points or lines, 2-D or 3-D) and the complexity of the estimation method. In recent work, several methods have been developed that provide closed form solutions for computing the transformation from the matched features.

The alignment approach avoids searching through the transformation space. However, it requires correct matching of initial features used to estimate the transformation. If only low-level features, such as points or lines are used, it may not be possible to obtain unambiguous matches. One commonly used approach is to match all subsets of features in the model to all subsets of features derived from the image. Clearly, such computation can be very expensive (it is $O(n^m)$,where n is the total number of features and m is the number required to compute a transform). Use of higher level features, either groups of features or even better, surfaces and volumes, can greatly reduce the complexity by reducing the ambiguity. Groups of features have been used for estimating the pose of objects in indoor scenes (for example, see [13]) application in outdoor scenes is likely to be much more difficult.

4. MATCHING FOR DEPTH ESTIMATION

To extract 3-D structure from two or more images, we need to compute correspondences between points or features in the multiple images; good surveys of various approaches can be found in ([5],[6], [8]). One fundamental difference between this task and that of registering an image to a site model is that a single, global transformation is not applicable. Rather, the transformation from one image point to another is a function of the unknown height of the point. Thus, we can not use global matching methods but need to match points and features in smaller areas. Small area matches, on the other hand, are highly ambiguous. A pixel by itself can only be characterized by an intensity value; this value varies somewhat with the viewpoint and many pixels in an image will have similar intensities. Some context from the neighborhood needs to be utilized to disambiguate.

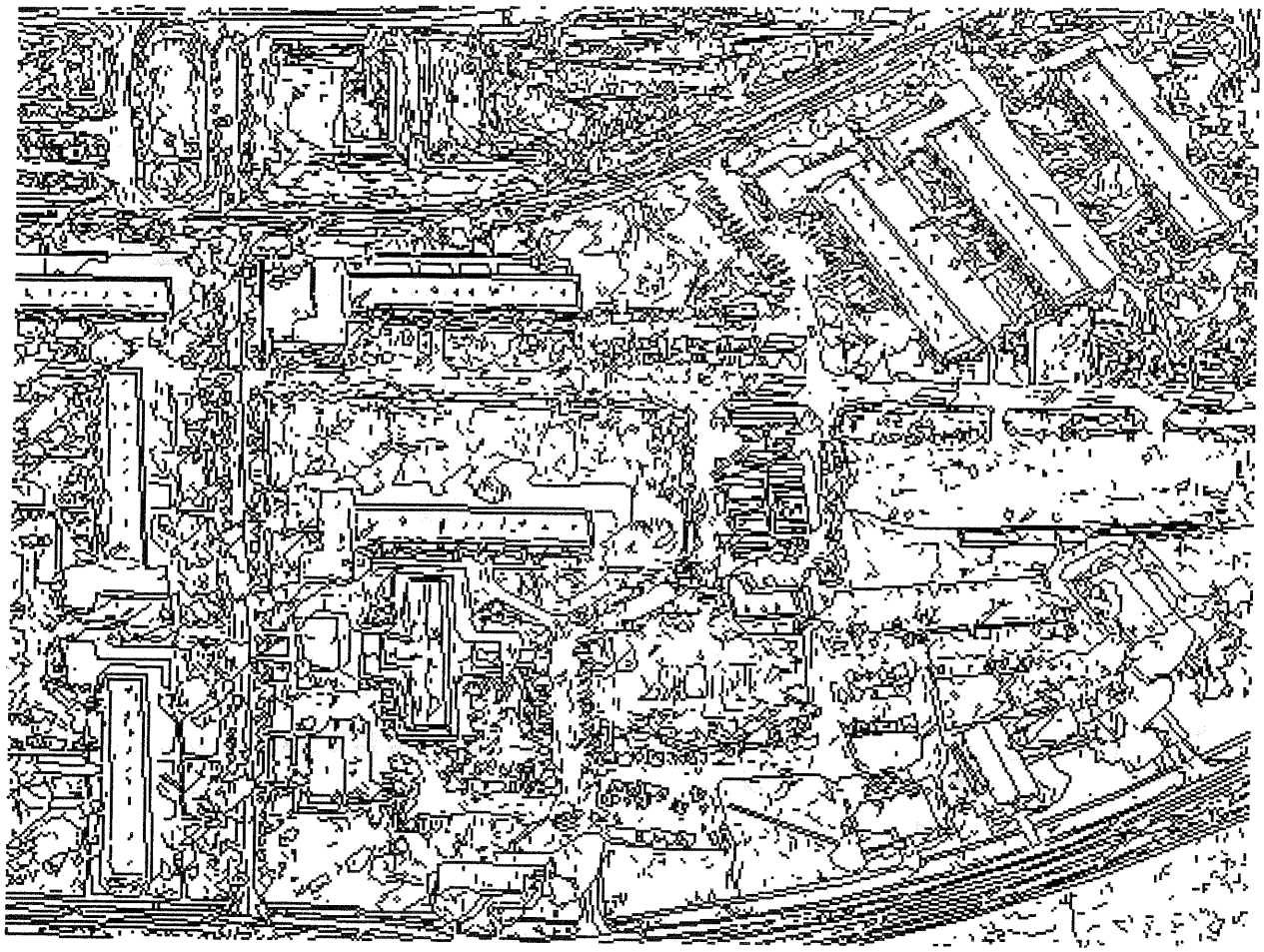

Figure 3 Line Segments

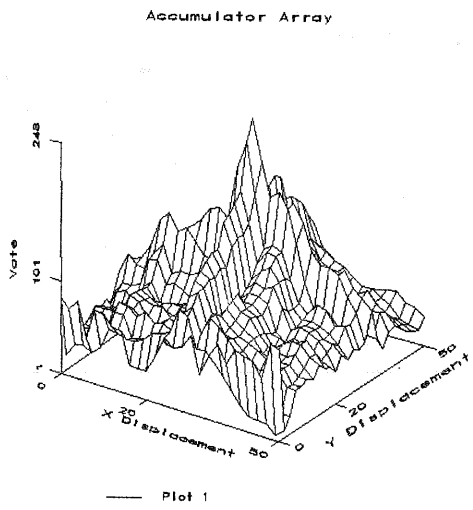

Figure 4 Accumulator Array

A good survey of the approaches to stereo matching covering methods upto the mid-80s may be found in [5]. Intensity correlation of small neighborhoods has been used for stereo matching since early times and continues to be popular even in current systems [9]. Such systems work reasonably well in presence of random texture and smoothly varying terrain, but are less effective in cultural environments with abrupt depth changes and large homogeneous areas (such as in scenes with many buildings).

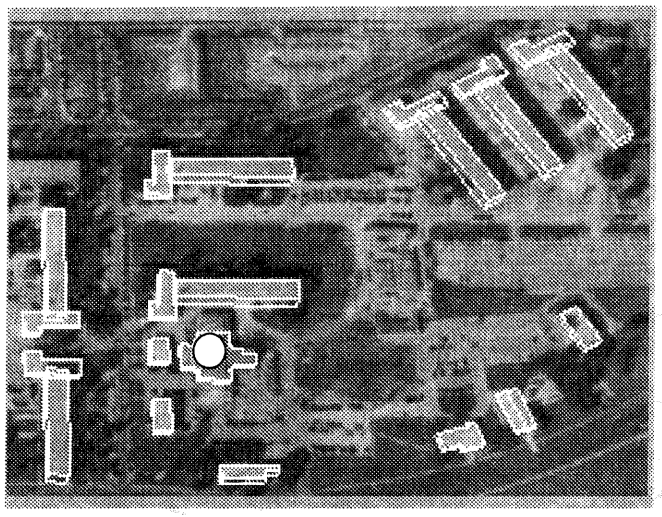

Figure 5 Registered Model

Feature matching techniques have been applied for scenes containing man-made objects as these objects tend to provide boundary features that are largely invariant to the viewing geometry (of course, they could be hidden). Individual features, such as a straight line, remain quite ambiguous. If more than two images are available, the ambiguity can be resolved by employing the epipolar constraints between the different pairs [1]. If only two images are available, matches of groups of features need to be considered to resolve this ambiguity and several techniques for this are described in the literature ([3], [4], [5], [8], [15], [18]).

Note that matching features such as lines necessarily gives only sparse depth information about the scene and detection of objects, such as buildings, requires a further step of grouping of lower level features. This leads to considering the construction of the desired groups first and matching the groups. A system following this approach for the detection of rectilinear buildings is described in [16]. In this system, parallelograms that may correspond to roof boundaries are hypothesized by grouping in the two images individually. These groups are then matched to select among them and to determine heights.

As observed before, there is a clear trade-off between the levels at which the stereo matching is performed. Higher level features can be matched with much less ambiguity, however, errors may be made in their computation in the first place. We believe that, in general, it is not possible to determine a correct level of matching in advance. Rather,

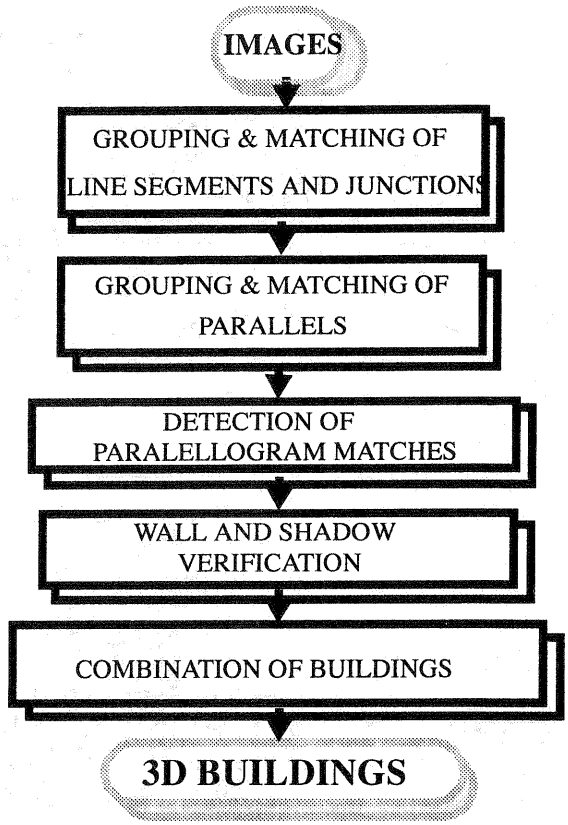

Figure 6 Block Diagram of Multi-View system

matching should occur at various levels, with the results of one level influencing the others and choice among matches made only when sufficient context becomes available to resolve confidently. An early system advocating this approach is described in [12]. We have recently constructed a system using such an approach [17] for the detection of rectilinear buildings: block diagram is shown in Figure 6. A characteristic of this system is that the images to be matched need not be taken at the same time; changes due to illumination and imaging conditions can be accommodated.

An example is shown in Figure 7 through 10. Figure 7 shows two views of a scene containing two buildings. Note that the two images are taken at different times and are of different resolutions. Figure 8 shows the lines that are detected in each image independently. The lines are then matched among the available images and sets of possible matches are computed. The matched lines are then grouped into higher level features (parallels and parallelograms) and the groups are then matched across the images (using the information about the matches of their constituents). Figure 9 shows the roof hypotheses.The hypotheses are verified by looking for looking for supporting evidence from the visible walls and the shadows. Figure 10 shows the two selected and verified buildings from the available hypotheses. At this point 3D models of the buildings have also been computed.

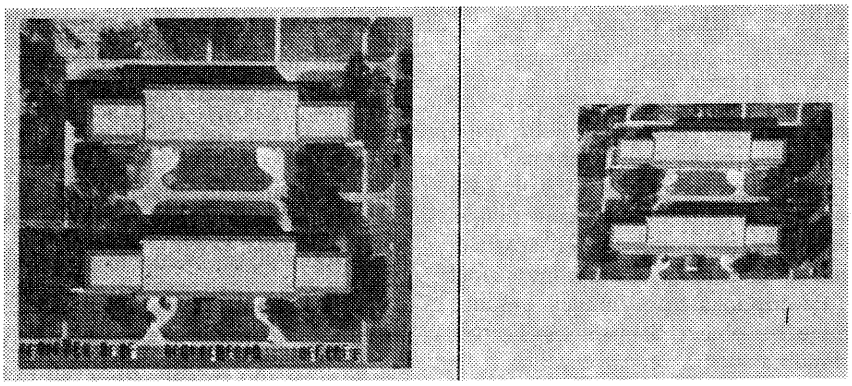

Figure 7 Two views of a scene

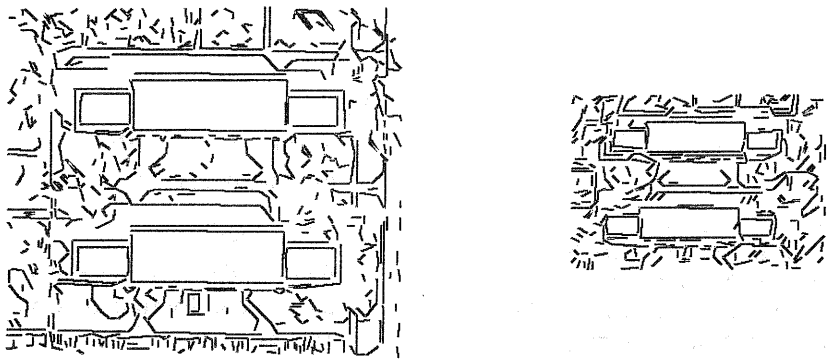

Figure 8 Lines detected from views in Figure 7

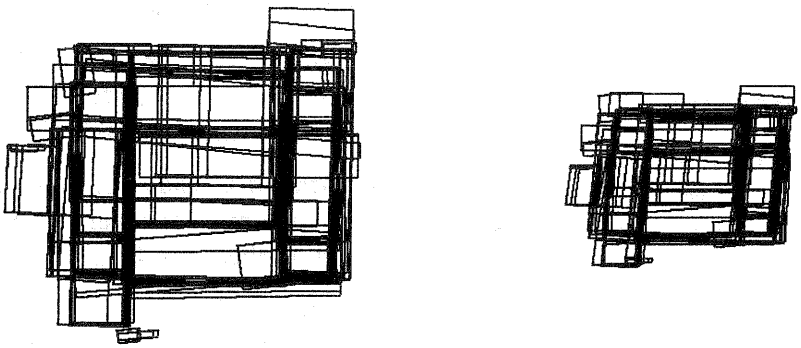

Figure 9 All roof hypotheses

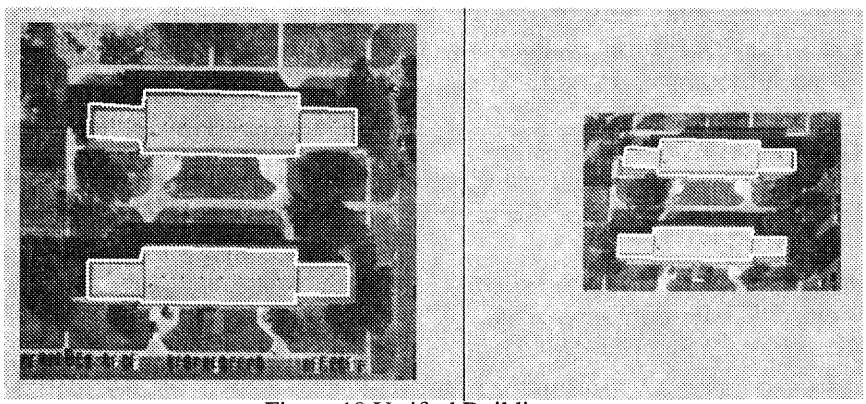

Figure 10 Verified Buildings

5. CONCLUSIONS

We have discussed the issues involved in choosing matching techniques for different tasks and illustrated with some selected examples. Matching of images with other images or with maps and models in complex environments remains a difficult and challenging task. It is this author's view that as the scene complexity increases, the matching problem can not be adequately solved without using more context and computing higher level descriptions from images, perhaps upto the point of finding the 3-D objects themselves or larger parts of their surfaces. Finding such objects, of course, is a difficult problem in itself and requires major advances in the techniques for perceptual grouping and scene segmentation. Fortunately, it appears that, in many cases, the segmentation and grouping processes can cooperate with the matching processes, reducing the complexity of both.

6. ACKNOWLEDGEMENTS

I would like to thank Prof. Heinrich Ebner for inviting me to present this paper. Andres Huertas and Sanjay Noronha have helped me in preparation of this paper and provided the examples used herein. The research reported here was supported by the Image Understanding program of the Defense Advanced Research Projects Agency (DARPA) under grant number F49620-95-1-0457 and contract number DACA76-93-C-0014 monitored by the Air Force Office of Scientific Research and by the Topographic Engineering Center respectively.

REFERENCES

[1] Ayache, N. J., and Lustman, F., 1991, Trinocular stereo vision for robotics", Transactions of IEEE on Pattern Analysis and Machine Intelligence, 13(1), pp. 73-85.

[2] Bejanin, M., Huertas, A., Medioni, G., and Nevatia, R., 1994, Model validation for change detection", Proceedings of Worshop on Applications of Computer Vision, Sarasota Florida, pp. 160-167.

[3] Boyer, K. L., and Kak, A. C., 1988, Structural stereopsis for 3-d vision, IEEE Transactions on Pattern Analysis and Machine Intelligence, 10, pp. 144-166.

[4] Chung, R. C. K. and Nevatia, R., 1992, Recovering building structures from stereo, IEEE Proceedings of Workshop on Applications of Computer Vision, pp. 64-73.

[5] Dhond, U. R., and Aggarwal, J. K., 1989, Structure from stereo - a review, IEEE Transactions on Systems, Man and Cybernetics, 19(5), pp. 1489-1510.

[6] Ebner, H., Heipke, C. and Holm, M., 1993, Global image matching and surface reconstruction in object space using aerial images, in Barrett E. B., McKeown D. M. (Eds.), Integrating photogrammetric techniques with scene analysis and machine vision, Proceedings of SPIE, pp.44-57.

[7] Forstner, W., 1982, On the geometric precision of digital correlation, International Archives for Photogrammetry and Remote Sensing, 24(3), pp.176-189.

[8] Grewe, L. L. and Kak, A. C. 1994, Stereo vision, in Chien, Y. T. (Eds.) Handbook of Pattern Recognition and Image Processing: Computer Vision, pp. 239-317.

[9] Hannah, M. J. 1989, A system for digital stereo matching, PE & RS, 55(2), pp. 1765-1770.

[10] Heipke, C. 1996, Overview of image matching techniques, OEFPE workshop on applications of digital photogrammetric workstations, Lausanne.

[11] Huttenlocher, D. P. and Ullman, S., 1990, Recognizing solid objects by alignment with an image, International Journal of Computer Vision, 5(2), pp. 195-212.

[12] Lim, H. S. and Binford, T. O., 1987, Stereo correspondence: a hierarchical approach, Proceedings of Image Understanding Workshop, pp 234-241.

[13] Lowe, D. G., 1987, Three-dimensional object recognition from single two-dimensional images", Artificial Intelligence, 31(3), pp. 355-395.

[14] Medioni, G. and Huertas, A., 1991, Automatic registration of color separation films, Machine Vision and Applications 4, pp. 33-51.

[15] Medioni, G. and Nevatia, R., 1985, Segment-based stereo matching, Computer Vision Graphics and Image Processing, 31(1), pp. 2-18.

[16] Mohan R. and Nevatia R., "Using perceptual organization to extract 3-D structures", IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(11), 1121-1139, Nov 1989.

[17] Noronha, S. and Nevatia, R., 1996, Detection and description of buildings from multiple aerial images, Proceedings of Image Understanding Workshop, Palm Springs CA, pp. 469-478.

[18] Roux, M. and McKeown, D. M., 1994, Feature matching for building extraction from multiple views, IEEE Proceedings of Computer Vision and Pattern Recognition, pp. 46-53.