On recognition of object categories

Pavel Grossmann*
Long Range Research Laboratory
GEC Hirst Research Centre
East Lane
Wembley
Middlesex HA9 7PP

A high level representation of polyhedral scenes in terms of planes and corresponding coplanar sets of 3D line segments is used to develop a method for identifying categories of objects and features in the scenes. Plane intersections are used to establish links between all the planes that correspond to the visible surface of a particular object. The object's shape (as far as it is known) is then reconstructed to provide its description and also constraints on its possible interpretations. At the same time the segment distribution within each plane is analyzed to search for any characteristic patterns that may help identification. In this way we use the topology of a 3D shape or a 2D segment pattern to identify a category of an object (a desk) or a feature (a window), rather that using a metric description of particular object or feature to find its instance(s) in the scene.

1 Introduction

The task of object recognition and scene interpretation is a challenging one. It embraces a large number of capabilities and methods ranging from simple template matching to the use of mathematical logic and extensive prior knowledge of the relevant domain in interpretation of previously unseen objects and scenes. The choice of methods and techniques used in a particular case depends very much on the nature of the task and on the available data.

The approach that we describe in this paper *(COMPACT)* is no exception. As part of the ESPRIT project P940 - *Depth and motion analysis*, we are developing recognition capabilities for a mobile robot operating within a man-made (indoor) environment (i.e. mostly polyhedral scenes) and also for a robot arm manipulating simple manufactured parts (whose surfaces comprise planes and simple quadrics). The input for the higher level processing modules is a set of 3D straight line segments produced by the three camera stereo vision system developed at INRIA (Rocquencourt, France) [1] for the project.

The first stage of *COMPACT* creates an intermediate representation of the image data in terms of planes, spheres, cones and cylinders. While these surface types are clearly sufficient to describe most indoor scenes, they also provide adequate description for many classes of manufactured objects.

A very strong incentive for creating a surface-based representation comes not only from the human visual experience but also from our previous choice of the 3D segment representation. Here, as we can see on the example shown in Figure 8, many of the vertices or junctions are missing as no special effort is made to find them in images and to preserve them during the stereo matching process. Hence the popular approach of "interpretation of line drawings" (see e.g. [2, 3, 4]) would be not very useful. One property however, that the disconnected segments corresponding e.g. to the windows in the scene still possess (and that can be extracted from the data) is their planarity. Our methods for extracting planes and other simple surfaces from line segment data have already been described elsewhere [5, 6, 7, 8].

The following stage, to which the rest of the paper is devoted, concerns the two aspects of recognition and interpretation that we can investigate in parallel using our representation - a 3D analysis of object shapes using the extracted surfaces [9] and a 2D analysis of surface features using the line segment distribution within each surface. Here we shall restrict our discussion to polyhedral objects and scenes for which we can already present some results.

2 Reconstruction of objects and spaces

A set of surfaces extracted from the image data constitutes a surface-based representation of objects or scenes that is clearly suitable for the recognition of known objects or scenes by direct matching to geometric models stored in the data base using any of the existing methods (e.g. [10, 11]). The small number of image features involved here reduces the time needed for matching.

In this paper, however, we investigate a different route to scene interpretation - via construction of 3D shapes from the available surfaces (3D scene segmentation) followed by a labelling stage in which geometric and topological *relations* between objects and other *constraints* are used to identify the shapes as object *categories* (rather

*This research was supported in part by the ESPRIT project P940.

AVC 1989 doi:10.5244/C.3.1

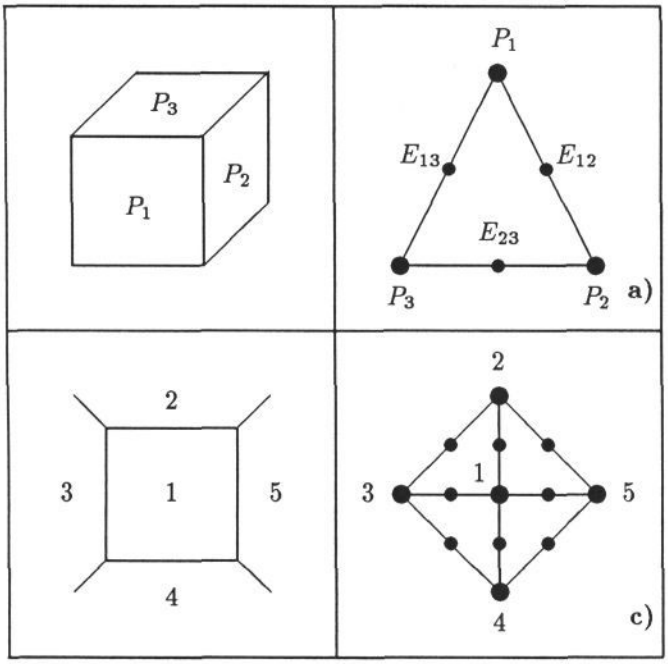

Figure 1: Connectivity of box surfaces
(the nodes correspond to planes P and edges E)
a) an object-box
b) a room-box

than particular instances) by assigning *labels* to them.

Our motivation is, having considered a range of applications corresponding to different domains and tasks (from object recognition to navigation), to create an *application neutral* representation in terms of 3D object- and space-primitives based on topology rather than metric description.

2.1 Basic ideas

We shall first consider the reconstruction of objects and spaces (in what is basically the *surface boundary representation*) in the domain of simple rectangular blocks or *boxes*. In a typical scene several *Object-boxes* are contained inside a *Room-box*.

The main structural component of a *box* will be an *edge* - an intersection between two visible planes. The simplest *Object-boxes* are obviously constructed out of convex *edges* and *Rooms* out of concave *edges*.

A *box* is constructed using the *connectivity* of its *edges* and planes which has been made explicit in our data structure. Starting with any *edge* we find the planes associated with it and in each plane we find other *edges* etc. This can be represented by a graph as shown in Figure 1. The task of constructing an *Object* or a *Room* can be then described as the task of finding *maximal complete subgraphs* of a graph containing all the planes in the scene.

Although the use of metric quantities like *perpendicularity* may provide convenient constraints in some domains or applications, they are not essential in this mainly topological representation and our approach is easily extended to general polyhedral scenes.

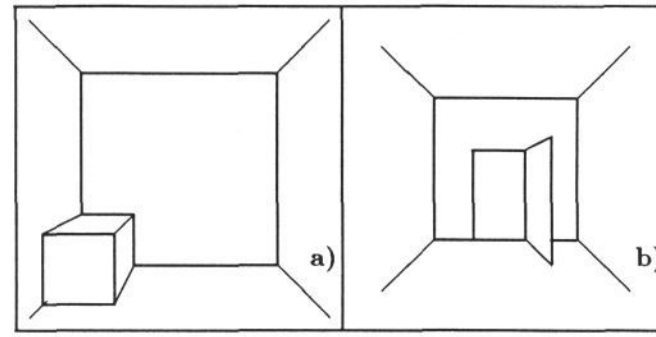

Figure 2: Complex scenes
a) a box in a room
b) a room with an open door

2.2 An *Object* in a *Room*

We shall now use this, the simplest nontrivial scene, to explain the basic *box*-building procedure that will be later modified to cope with scenes of increasing complexity and eventually with the general case of composite objects and spaces.

An *Object* in a *Room* (Figure 2a) will give rise to some concave *joins* between the *Object* faces and the *Room* floor or walls that should not be confused with the *Room constituent edges*. Hence the *box*-building procedure will involve the following steps :

1. Label all *edges* as convex or concave.
2. Use the convex *edges* to construct the *Object*.
3. Label any concave *edge* that is associated with the *Object* as a *join*.
4. Use the concave *edges* that are not *joins* to construct the *Room*.

2.3 Complex scenes

A group of desks in an office is an everyday example of a scene where several *object-boxes* may share a plane. Our basic procedure has to be modified to distinguish between different physical surfaces (i.e. desk-tops) in the same plane. Here we make use of the relative proximity of edges without actually requiring the existence of 3-edge vertices.

Another example is a room with an open door (Figure 2b) seen as a lamina (rather than a *box*). The door plane, unlike the other vertical planes (walls), usually divides the scene into two halfspaces, both of which are (at least in part) visible to the camera. This can be used to identify it as being different from the walls. An alternative method which uses the relative position of the door-wall intersection with respect to the wall-wall intersections is currently being tested.

2.4 Composite Objects

In order to extend our method to composite objects (in our case unions of several convex parts created in the spirit of the *constructive solid geometry representation*)

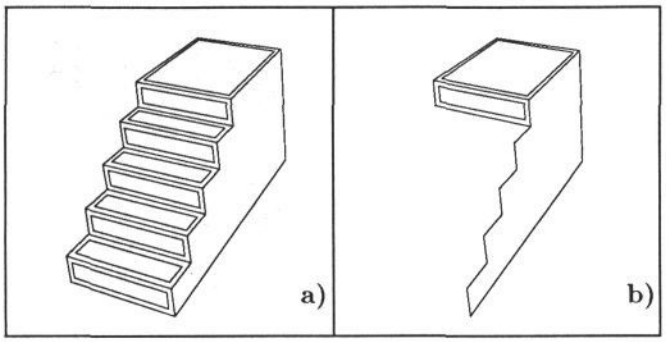

Figure 3: Composite object - a staircase
a) the whole object
b) one of the five convex parts

we have to adopt an operational definition of a single object. As it stands now, our method will identify as a *box* every *box*-like shape in the scene - be it a simple object or a convex part of a composite object. Such elementary *box*-parts have to be merged to create meaningful single objects.

The operational definition necessarily depends on the application domain. Our box-world environment is relatively simple and so one might expect a relatively simple definition. On the other hand our surface representation does not offer the usual clues to the integrity (or otherwise) of an object - e.g. colour or texture. At this stage we also assume no higher level knowledge regarding the possible function of an object to guide us.

Let us consider a pair of *box-parts*. First we require that the two *box-parts* have a common plane. Then we look for an evidence that they are actually joined together. For example, the staircase in Figure 3a is first reconstructed as a set of five *boxes* corresponding to individual stairs (Figure 3b). Then the pairs of adjacent *boxes* are labelled as *connected* becaused they have a common plane and also a visible (concave) *join*. Inevitably all the *boxes* are identified as parts of the same object using these two requirements.

The integrity of the common planar surface itself may also indicate that parts are connected (at present we assume that all surfaces are opaque). Preliminary tests with different synthetic objects and scenes produced a variety of indicators and rules which are currently being put on a sounder theoretical basis.

2.5 Shape interpretation

While some objects have a simple box-like shape (e.g. a filing cabinet) and their identification requires some additional information, in some cases (a desk) the shape alone may provide a constraint sufficient for identification.

From the set of segments representing the desk in Figure 4 our program extracts two *boxes* : box A comprising planes P1,P2 and P3 and box B planes P3 and P4. It is important to note that these *boxes*, because of the way they are constructed, cannot be simply identified with the usual block-parts we may use to build such a shape. Hence our shape description will differ from the usual schemes (a horizontal block supported by two upright ones, see e.g. [12]). All we can say is, that the two boxes

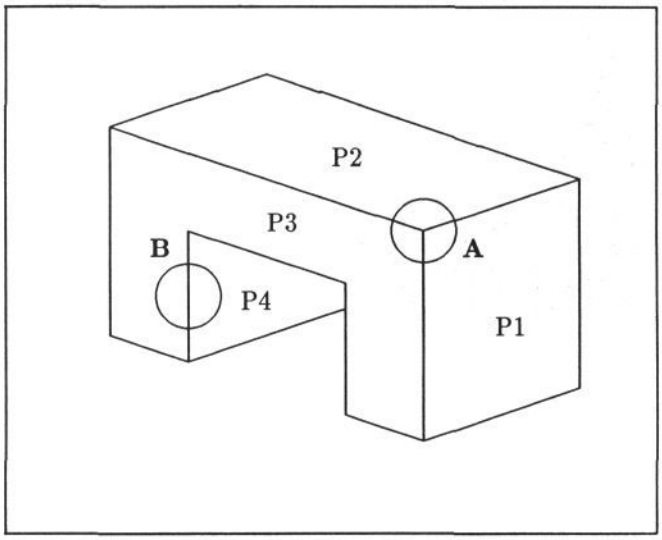

Figure 4: A desk

share a plane (P3) and that the planes P1 and P4 are parallel; furthermore box B is in a way "contained" by box A. While this does not amount to a recognizable description of a desk (or an arch), it may enable us to choose one from a small number of interpretations.

Here we adopt an approach similar to that in Minsky's frame representation [13]. We consider several basic types of scene (e.g. an office) and in each scene we expect to find a small number of objects and features (desk, chair, wall, window ...). When interpreting a particular 3D shape, we need not consider the domain of all possible (polyhedral) objects, only a few.

Furthermore, some objects may be expected to possess, apart from a particular shape, a characteristic surface pattern of lines - e.g. indicating a set of drawers in a desk. In such a case the interpretation module can initiate a relevant analysis of the edge segment distribution to search for such a pattern (see next section).

2.6 A simple example

While the synthetic staircase in Figure 3 nicely illustrates our basic method of object reconstruction, the real data in Figure 5 emphasizes the importance of making explicit the planes and their intersections (rather than vertices) that is fundamental to our approach. The set of 3D segments from ITMI (Grenoble, France) that represents the polyhedral object in the scene (electrical switch) was extracted from a series of images taken by a single camera mounted on a robot arm moving around the object.

Although the data is quite sparse, we succeeded in identifying five planes and establishing their connectivity. Our (incomplete) knowledge of the shape, represented in Figure 5d, would be adequate for simple classification and grasping tasks. (To visualize our results, we find the smallest rectangular block that contains all the segments and use the identified planes to cut away the parts that are outside the object. Hence we are always dealing with a finite volume. The *back* and *bottom* faces in Figure 5d are parts of the block surface.)

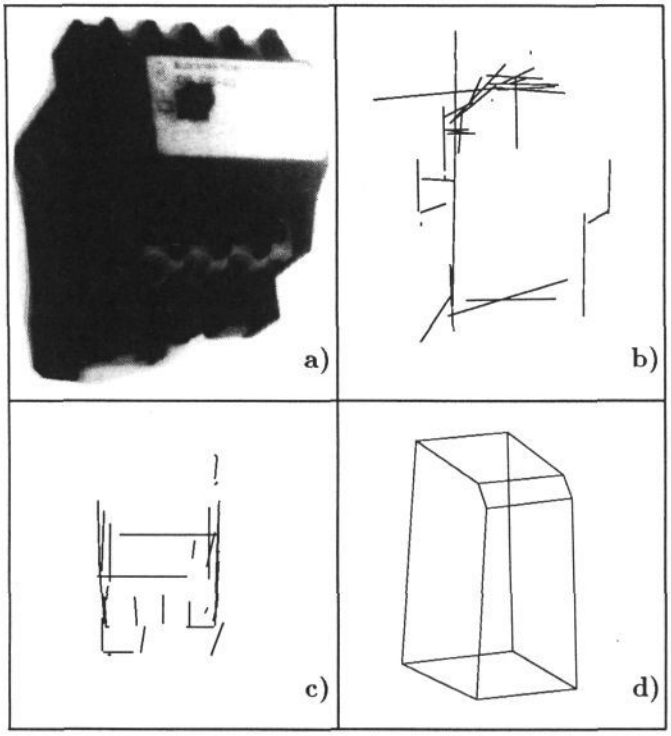

Figure 5: ITMI polygonal object
a) one of the series of images
b) side view of the 3D segments
c) top view
d) reconstructed shape

3 Planar segment patterns

3.1 The pattern primitive

In the 3D shape reconstruction process we used only the knowledge of the plane parameters and the plane segments that are part of the plane intersections. Very often, however, the other segments in the plane can provide valuable clues as to the nature of the corresponding surface and hence of the relevant object or space. For example the front face of a box-like object may contain a pattern corresponding to a set of drawers thus identifying the object as a filing cabinet; a window-like pattern may identify a wall in a room.

Our approach to such pattern analysis is again determined by the basic recognition of the fact that the line segment data available in practice is usually far from perfect, e.g. we expect many of the line junctions to be missing (e.g. Figure 8).

Although the obvious representation primitive for many characteristic patterns in man-made environments (e.g. a window) seems to be a rectangle, the missing line junctions can cause serious problems in the extraction of such primitives. So instead we chose a straight *ROW* of parallel segments (Figure 6) that are, like the rungs of a ladder, equal in length and perpendicular to the ladder axis. The distances between the adjacent segments (i.e. gaps), however, need not be all equal.

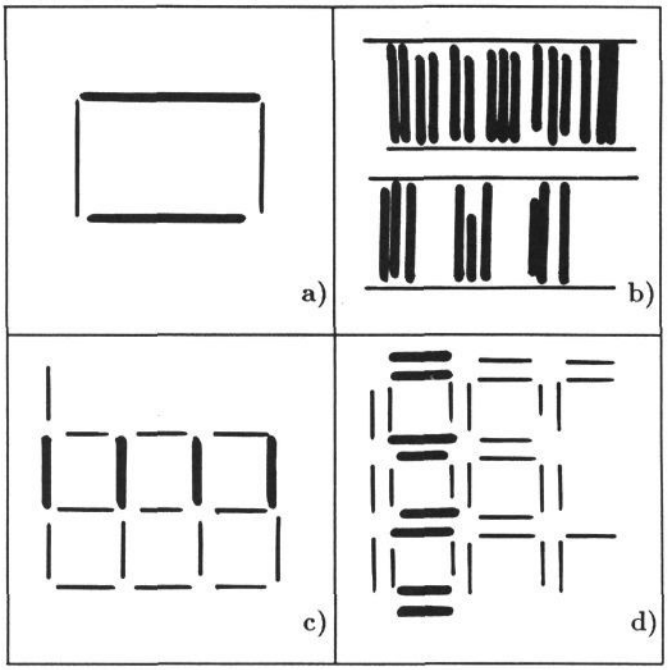

Figure 6: Types of rows and patterns
a) FRAME
b) BOOKS and DRAWERS
c) TILES
d) WINDOW

3.2 The pattern analysis

The whole process of pattern identification and interpretation can be separated into three stages. Firstly we find all the *rows* in the segment data to create the intermediate representation, and also compute both the individual row properties and the row correlations.

Then we group these primitives to create certain distinctive patterns or *topological* features like *WINDOW* (Figure 6). As the segment data is rather sparse, we decided against the statistical approach to texture analysis as advocated e.g. by Vilnrotter [14].

Finally we have to interpret these patterns in terms of real object features. This process is very much domain dependent and we have to use sets of a priori probabilities linking the physical and the topological features. As an example a window in the scene is most likely to give rise to the pattern *WINDOW*, but also *TILES* or *FRAME* (Figure 6) are possible.

3.3 Row properties

For every *row* we can determine its individual properties like its orientation (given by that of the segments), *size* (number of segments), *gnum* (number of different gap widths) *rank* (the largest number of equal gaps), as well as a *regularity index RI* (which is somewhat similar to entropy in physics) defined as :

$$RI = \frac{size - 1}{size - 2}\left(1 - \frac{gnum}{size - 1}\right)$$

so that its values range from 1.0 (all gaps the same size) to 0.0 (all gaps different sizes).

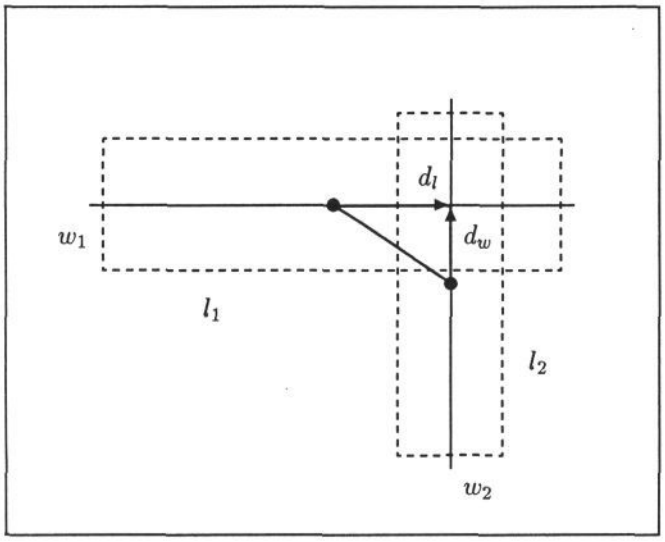

Figure 7: Two overlapping rows

We also make explicit the relative orientation for each pair of *rows* if they are parallel or perpendicular. Such correlations are used to find groups of *rows* likely to belong to the same topological feature. Here we require two *rows* not only to be perpendicular but also to overlap. The definition of *row* overlap is given in Appendix A and Figure 7.

Using a whole set of individual properties and pair correlations we at present define and distinguish several characteristic row types : *FRAME, DRAWERS, BOOKS, WINDOW* and *TILES* (Appendix B and Figure 6). As the names suggest, these row types are expected to indicate presence of real features like windows, tile patterns or rows of books.

3.4 Topological patterns

The next step is to group together *rows* likely to belong to the same topological pattern. We make an observation that two perpendicular rows belong to the same pattern if they overlap. We call such rows *complementary*. Starting with a *row* R of a particular type T we proceed to include all rows *complementary* to R and every *row* of the type T parallel with R that shares with R at least one *complementary row*. This procedure establishes just the right degree of connectivity for sensible grouping. The resultant pattern is defined to be of the type T.

This method of pattern analysis was applied to the set of 3D segments representing an office scene at INRIA shown in Figure 8a. In Figure 8b we show the subset of segments identified with the feature *WINDOW*.

3.5 Feature interpretation

As mentioned earlier, we do not expect simple one-to-one correspondences between the topological patterns and the physical features; a window in the scene may give rise to a *FRAME*, a *WINDOW* or even a *TILES* pattern and on the other hand a *FRAME* can correspond to anything from a doorway to a notice on the wall.

Each type of scene will require an interface that, using some domain specific data, will assign interpretation probabilities to each segment pattern. Hence, for each characteristic pattern found in the segment data we get, as the result of this interpretation stage, a set or possible interpretations with the corresponding probabilities.

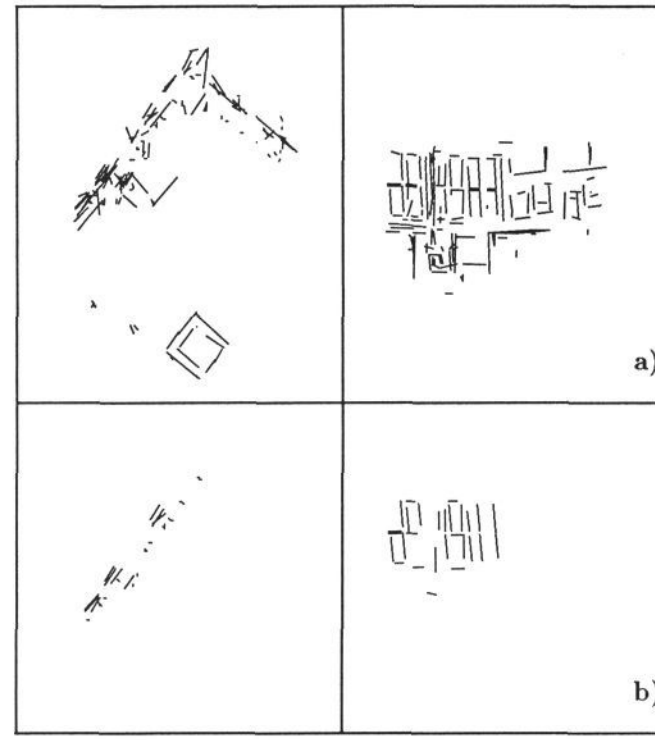

Figure 8: INRIA office scene
a) top and front view of the 3D segments
b) the WINDOW pattern

4 Summary

We have outlined a method for object recognition and scene interpretation which is strongly surface-based. Surface primitives are extracted from the 3D line segments directly without referring to the surface contours. They are combined to form objects in 3D using connectivity via surface intersections that are made explicit by the method, again without the necessity to detect or reconstruct corners and vertices.

Within each surface (plane) we search for characteristic patterns of line segments that may help to identify the object or feature. The results obtained so far justify a certain amount of optimism, although more extensive tests on real complex scenes are still required.

At present we are investigating the *topological descriptions* of the 3D structures suitable for recognition (e.g. desk-like or staircase-like) and also the combined use of the 3D shape and the surface patterns in object and scene interpretation.

5 Acknowledgements

Thanks are due to my colleagues in the ESPRIT P940 collaboration (INRIA at Rocquencourt and Sophia-Antipolis and ITMI in Grenoble) for the use of their

image data and to Mike Brady for many inspiring discussions.

A The row overlap

In order to define an overlap of two perpendicular rows we represent each row by a rectangle specified by the row's centre point C, its orientation and its length $2l$ and width $2w$ as indicated in Figure 7. Two rows are deemed to overlap if their corresponding rectangles overlap. Let $\vec{d}$ be the vector connecting the two rectangle centres and let d_l and d_w be its two components along the length and width of row 1. The absolute overlap condition is :

$$d_l < l_1 + w_2$$

$$d_w < l_2 + w_1$$

Maximum overlap is achieved when :

$$d_l + w_2 < l_1$$

$$d_w + w_1 < l_2$$

In practice we use the following simpler condition :

$$d_l < l_1$$

$$d_w < l_2$$

B The characteristic row types

Pairs of perpendicular overlapping rows with $size = 2$ form a separate class. They are simple rectangles and in our analysis they are given the label FRAME.

Otherwise we aim to characterize only the larger ($size \geq 4$) regular ($rank \geq 3$) rows. Those perpendicular to other large ($rank \geq 2$) rows suggest "extended" features and are labelled as TILES if they are highly regular ($RI > RI_{th}$) and as WINDOW otherwise. RI_{th} is a threshold value that depends on the row size (the maximum for WINDOW type).

Rows perpendicular to smaller or less regular rows ($rank = 1$) are in the "linear" category : the regular ones ($RI > 0.3$) are DRAWERS and the large ($size > 10$) less regular ones ($RI < 0.3$) are labelled as BOOKS.

These assignments are specific to the "office scene" domain and preliminary.

References

[1] Nicolas Ayache and Francis Lustman. Fast and reliable passive trinocular stereovision. In *Proceedings of the First International Conference on Computer Vision*, 1987.

[2] D.H.Ballard and C.M.Brown. *Computer Vision.* Prentice Hall, Inc., 1982.

[3] Martin Herman and Takeo Kanade. Incremental reconstruction of 3D scenes from multiple, complex images. *Artificial Intelligence*, 30, 1986.

[4] Kashipati Rao and R. Nevatia. Generalized cone descriptions from sparse 3-d data. In *Proceedings CVPR'86*, 1986.

[5] Pavel Grossmann. COMPACT - a 3D shape representation scheme for polyhedral scenes. In *Proceedings of the Third Alvey Vision Conference*, 1987.

[6] Pavel Grossmann. Building planar surfaces from raw data. Technical Report R4.1.2, ESPRIT Project P940, 1987.

[7] Pavel Grossmann. COMPACT - A surface representation scheme. In *Proceedings of the Fourth Alvey Vision Conference*, 1988.

[8] Pavel Grossmann. Planes and quadrics from 3D segments. Technical Report R4.1.6, ESPRIT Project P940, 1988.

[9] Pavel Grossmann. From 3d line segments to object and spaces. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1989.

[10] O.D.Faugeras and M.Hebert. A 3d recognition and positioning algorithm using geometrical constraints between primitive surfaces. In *Proceedings of the Eighth IJCAI*, 1983.

[11] E.Grimson and T.Lozano-Perez. Model-based recognition and localization from sparse range or tactile data. *International Journal of Robotics Research*, 3, 1984.

[12] P.H.Winston. Learning structural descriptions from examples. In P.H.Winston, editor, *The psychology of computer vision.* McGraw-Hill Book Company, 1975.

[13] M.Minsky. A framework for representing knowledge. In P.H.Winston, editor, *The psychology of computer vision.* McGraw-Hill Book Company, 1975.

[14] F.M. D'Auria Vilnrotter. *Structural analysis of natural textures.* PhD thesis, University of Southern California, 1981.