RECOGNITION AND DEPTH PERCEPTION OF OBJECTS IN REAL WORLD SCENES*

Robert J. Douglass

Department of Computer Sciences
University of Wisconsin-Madison

Automatons and other robotic applications which are designed to move around and interact with their physical environment need a computer vision system for recognizing and understanding the spatial relationships of objects in real world scenes. The perceptual system must be able to identify salient objects in a scene, develop an understanding of their spatial relationships, and maintain continuity from one view to the next as either the objects or the system's camera moves through the scene.

Outlined here and described in more detail in Douglass, 1977, is a system which has been implemented in SIMULA and tested on hand coded outdoor scenes of simple subjects such as houses and automobiles. It uses a recognition cone, a segmentation algorithm for dividing a scene into similar regions and a routine for constructing a three dimensional world model. Visual inference routines interpret perspective, shadows, highlights, occlusions, shading and texture gradients, and monocular motion parallax. Other visual knowledge is added with long term models and short term object representations.

The final program will be tested on color photographs of outdoor scenes using as input a series of views of the same scene from different angles which approximates what an automaton would "see" as it moves down a street.

Scene Description Program

The program's recognition cone (after Uhr: 1972) is a parallel-serial cone structure consisting of a number of processing layers. The first layer contains an array of light intensities in three colors as digitized from the system's camera. Successive layers average the picture, compute several measures of texture and color, and detect edges and angles. In the higher layers of the cone, object names are assigned to various areas based on the presence of certain configurations of edges and lower level features. The last layer contains a number pf textural, color, and edge descriptors for each point of the array and one or more possible interpretations for that point. '

The output of the cone is segmented into regions by a region growing algorithm similar to Yakimovsky's, which uses a Zobrist-Thompson grouping operator to estimate the probability of an edge between two points in the picture array. The region grower builds a description of each segment as it is grown including texture, color, size, adjacent segments, order of connectivity, brightness, and a list of possible interpretations for that segment.

The segments produced by the region grower are built into a three dimensional world model by a placement routine, the heart of the whole system. This routine uses the descriptions of the segments, a set of long term models of objects, the visual inference routines, and any contents of the world model from the previous view to form the segments into three dimensional surfaces representing objects in the scene. The routine begins by making an initial approximation to the segments depth or orientation in space and by selecting the highest weighted object name from the list of possible interpretations produced by the recognition cone. Each inference routine then successively improves the placement *of* the segment by using information in the long term object models, comparing the current placement with the previous contents of the short term memory, or examining the placement and interpretation of neighboring segments.

The visual inference routines assist in interpreting the objects'positions in the scene with general information about the relationship between the visual properties of the physical surfaces in the real world and the manifestations of these properties on the program's digitized input array. The routines contain heuristics for a) deciding when one segment is occluding another, in shadow, or highlighted, and b) interpreting shading and texture gradients, linear perspective, and motion parallax due to the movement of the camera between successive views of the scene.

Long term object models are used by both the visual inference routines and the placement routine. They provide a general description of shape, size, orientation, and expected context for the object within a scene and contain weights for each part of the description indicating the system's confidence. For example, if the occlusion routine, part of the visual inference routine, finds a segment labeled window adjacent to a segment labeled wall, it will consult the object model for a window and find the window segment has a high probability of touching the wall segment and is therefore lying in the same plane rather than occluding it.

The system integrates a sequence of slowly changing views of one scene over time by using a motion parallax routine. Segments from one view of the scene are matched with segments from the next view to compute their shift. This shift and the parameters describing the camera's movement between the two views is used to compute the orientation of a segment in space.

*This work is partially supported by the National Science Foundation, NSF Grant MCS76-07333

References

Douglass, Robert "Recognition and Spatial Organization of Objects in Natural Scenes." Computer Science Dept. Tech. Rept., Univ. of Wise, 1977.

Uhr, L., "Layered 'Recognition Cone' Networks that Preprocess, Classify, and Describe." IEEE Trans. Computers, 1972, 21, pp. 758-768.