RESOLVING VISUAL AMBIGUITY WITH A PROBE

VISION FLASH 17

by

John Gaschnig

Massachusetts Institute of Technology

Artificial Intelligence Laboratory

Vision Group

July 1971

ABSTRACT

The eye-hand robot at the Artificial Intelligence Laboratory now possesses the ability to occasionally copy simple configurations of blocks, using spare parts about whose presence it knows. One problem with which it cannot cope well is that of ambiguous scenes. This paper studies two types of ambiguity present in some scenes -- occlusion and illusion -- and proposes some ideas about effectively resolving the ambiguities through the use of the hand as an information detection device to work in conjunction with the eye.

Work reported herein was conducted at the Artificial Intelligence Laboratory, a Massachusetts Institute of Technology research program supported by the Advanced Research Projects Agency of the Department of Defense, and was monitored by the Office of Naval Research under Contract Number N00014-70-A-0362-0002.

Table of Contents

1. Introduction

The eye-hand robot at the artificial intelligence laboratory now possesses the ability to occasionally copy simple configurations of objects, using spare parts about whose presence it knows. An interesting problem with which the present robot system cannot well cope is that of ambiguous scenes. This occurs when there is not enough information present in the two-dimensional line drawing of the objects in the scene to completely characterize those objects. The problem may be one of identification of of the types of certain objects (i.e., whether the object is a block or a wedge), or it may be one of determining the dimensions of certain objects or their locations in space. This paper studies two types of ambiguity present in some scenes--occlusion and illusion--and proposes some ideas about effectively resolving the ambiguities.

First of all, the problem is a heterarchical one. Consider the drawing in fig. A-10 (in the appendix). The drawing presents an optical illusion, in that it can be interpreted either as a wedge resting on a block (house-shaped figure) or as a wedge abutting and partially occluding a block. The body finder module may find one or the other of the two models (depending upon one's luck that day, how well the preceding modules have done their jobs, and upon which of the various flavors of body-finding heuristics are in use), and pass on its answer. However, higher level modules have no way of knowing whether the interpretation proposed is correct or if there are other possible interpretations. They make their arm movement plans using the information

available and set the arm in motion to do its job. Only if the arm sends back a signal that it was unable to perform the actions requested of it that the plan-making modules may suspect that they were deceived by the body finder module.

In order for the robot to carry out the requests made of it when the visual scene contains ambiguities, there must be some way for the robot to interact with the real world in order to resolve the ambiguities. What is proposed here is to provide the robot with the ability to disambiguate scenes by developing and executing a plan of arm movements which will cause the arm to touch certain objects, test for their presence or absence in particular locations, and perform other such actions to determine empirically the state of the universe.

> Heinrich A. Ernst |1| created a robot system ten years ago which performed this type of interaction with its environment. His robot had no vision, but determined the state of the universe through sense devices on the hand of a mechanical arm which was moved around a table top strewn with blocks and a box. He wrote several programs in a goal-oriented language which executed searching and manipulative activities using the arm. In his thesis, Ernst raised may of the important questions underlying the present development of cognitive robot systems.

2. Description of ambiguities of occlusion and illusion

A few general observations about occlusion and illusion are in order at this point. A group of objects in a scene is ambiguous by occlusion if one or more objects are (from the point of view of the camera) in front of, above, or otherwise partially occluding one or more other objects, so that the entire outline of the occluded object(s) is not in the visual field. Thus, there may not be enough information present in the visual field to correctly identify the occluded object, or to establish its dimensions, or to determine its location in real space.

Fig. A-6 illustrates the case in which the identity of the occluded object cannot be determined from the visual information. The bottom object can be either a wedge or a block.

Fig. A-4 displays the case in which the dimensions of an occluded object, in this case its height, cannot be determined.

A group of objects in a scene is ambiguous by illusion if two or more objects marry each other whose identities are undetermined, in the sense that there are two (or possibly more) interpretations of the identities of the objects. The drawing in fig. A-10 is ambiguous by illusion.

3. Report from the body finder

Presumably, if the system is to resolve ambiguous scenes, one of the modules involved in the scene analysis must take on the responsibility of reporting (to a higher level module) that certain bodies in the scene under consideration are ambiguous. The body finder

seems to be a likely candidate for this job. It is at this point that there is just enough information to determine that a scene is ambiguous. Also, later modules should probably like to know the confidence level of the results of their predecesors. Rattner [2] makes some helpful suggestions in this respect. He proposes that the body finder generate a plausible interpretation of how the regions are connected, but also that it keep a record of alternatives, in case its favorite interpretation is rejected by higher level modules. It then should send the most plausible of its alternatives, and so on, until one is accepted.

If it becomes necessary for a higher level module to verify the body finder's model, it would be helpful for it to know if the scene is ambiguous or not. Thus it would be nice if the body finder could pass along this information also in the case of ambiguities, especially those of illusion, where there are usually only two possible alternatives.

4. Resolution of ambiguous scenes

4.1 Using the eye

Now we come to the problem of actually resolving the ambiguity. It is of course the case that the arm can be used to perform the disambiguation exercise, but this is not the only possibility. Stereo vision and depth perception through focus provide viable alternatives. Using stereo in the case of illuisions, the pictures from almost any two camera orientations should provide enough information to resolve the ambiguity. When the ambiguity is by occlusion, however, the position of

the camera is important and some care must be taken to insure that two views will be sufficient. Focus should work well in cases of illusion where the knowledge of the precise location of an ambiguous vertex will provide enough information to decide which interpretation is valid. This method will generally not be helpful in cases of occlusion.

Another possibility which may be feasible is to use both the eye and the arm to resolve the ambiguity, i.e., to have the arm perform some action and have the eye take a look at the result. In cases of occlusion, this may be very useful. For example, consider the drawing in fig. 1. The lying block rests in front of and occludes the standing block, so that the dimensions and the exact location of the latter are indeterminable. If the arm can remove the lying block, the eye can then look at the standing block unoccluded. Its dimensions are then easily found. In the case of illusions, the arm may be requested to move one of the objects, after which the eye can look at the result and determine the identities of the objects.

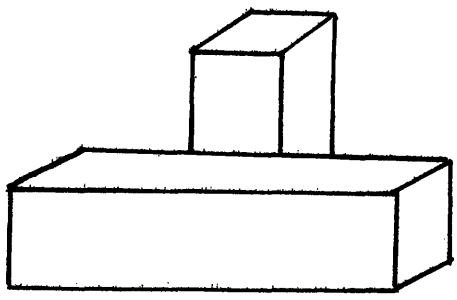

Fig. 1

In any case this process will be time-consuming. The proper action for the arm must be determined, its plan created and executed, another picture taken and processed, and more hand movements undertaken to restore the scene to its original form. If the objects in which the robot is interested happen to lie on the bottom of a stack, this type of procedure can become unwieldly. If the robot can assure itself that it will not need to look at the scene again, then perhaps the action of putting the removed objects back will not be necessary. But it does not appear that the confidence level of this assurance will be always high, since more, yet unrecognized ambiguities may lie ahead. Also, the movements of the arm must be very precise if it is indeed necessary to restore the scene to its original state. (The present arm does not posess this capacity in any sense.)

Another fundamental objection arises in the example of a T-shaped structure as in fig. A-4. In order to determine the height of the supporting block, it might be nice for the hand to pick up the block forming the crossbar of the "T", and then scan the now unoccluded supporting block to determine its dimensions. However, to pick up the supported block, the robot must know where it is. To calculate that it

must know the height of the supporting block, which fact it can't determine until it has picked up the supported block....a loop with no entry point.

It can be done. The robot can figure out where the supported block should lie in the xy-plane. It can then direct the arm to descend with one of its micro-switches in that column, until contact is made with the block. At that point it may be able to compute the height above the table of the top of the supported object and whence the height of the supporting object. Failing at that it might then, instead, pick up the supported object and take a new scan. The length of the above sequence suggests that it is not a very good way of doing it if many objects are involved.

4.2 Using only the hand

Using the arm without the use of the eye for resolution of ambiguities has the advantage that it does not require the lengthy scanning and processing of a visual image. The question then becomes: restricting the arm to touching objects but not allowing it to move them, what is the manner in which the arm should be used? to aid in this study an experiment was devised and conducted. The subjects were shown ambiguous scenes both of occlusion and illusion and asked how they would resolve the ambiguity in each scene, using only the index finger to touch the objects in the scene. They were told to assume that they had the ability to move their fingers accurately to any spot in space, although they were told they could not assume that they could watch what their

fingers were doing. The appendix describes the experiment in full. It would seem that in many cases (especially those of illusion) the easiest and most direct approach would be to place the finger in an appropriate spot and test whether that spot was occupied or not. However, all but one of the subjects distrusted this unfamiliar ability enough that they did not even consider using it, except to place the finger in an initial position on a surface of an object. Instead, each devised procedures suited to the particular case of moving his finger on the surfaces of the objects. Among the tests made to resolve the ambiguity were counting the number of sides of an object around its horizontal cross section (i.e., counting the number of changes in direction), measuring discontinuities in direction when moving along a surface, and the like. Such actions are generally more complex than merely using the finger as an "occupied" predicate, and they underline the importance of precision of movement of the arm if such activities as the latter are to be undertaken.

In cases of illusion it seems that the best approach is to use the micro-switches of the hand to test whether or not a space is occupied. The crux of the problem is to determine which point to send the hand to. In one possible arrangement the body finder sends a description of one of the plausible models to its successor programs, in addition to a flag telling that the correctness of this model is not particularly certain. The disambiguator program would not know what the other model looks like. Without this information it is still possible to verify the validity of the proposed model. Since we are dealing with planar objects, two objects having the same vertices implies that they are identical. So

if the hand is sent to test for the presence of the vertices in the places predicted by the proposed model and finds them all to be present, then the interpretation is indeed correct. Conversely, if any vertex is missing then the interpretation is false. This procedure is cumbersome, however. Its actions are more numerous than should be required. Furthermore, even assuming the availability of precise arm movement, the presence of vertices is difficult to verify with great accuracy.

If the disambiguator gets from the body finder the other likely alternative, things are much simpler. It can compute this model's supposed coordinates and calculate the section of real space that is occupied by objects in both interpretations. Then it can pick a point of space which is occupied in only one interpretation (probably by a weighting of such predicates as "easy for the hand to reach" and "far from any spot which is either occupied in both interpretations or in neither). If the hand approaches and passes this point without activating its micro-switches, then that space is unoccupied. If the micro switches indicate that the hand cannot reach the point, then it is occupied.

4.3 Hardware constraints

The use of the hand to move along surfaces and test for discontinuities, as suggested by the subjects of the experiment seems to have some merit, especially in cases of illusion, but, alas, with the present arm this is out of the question. This gliding along the surface of an object is a very complicated type of action, requiring a constant feedback loop to the software to direct the hand to be in constant

minimum-pressure contact with the surface. This feedback loop does not exist; indeed, the switches on the hand provide only a few very primitive interupt signals. Furthermore, the arm is mechanically incapable of such smooth movements. Its movements are jerky and an attempt to use the arm as described above would be liable to send the blocks sprawling across the table top. The new arm will likely also have this problem, as any movement in the x-y plane will be done by the overhead crane, whose slight discontinuities in smooth motion caused by changes in velocity are likely to be amplified in the hand. This method may be worthy of consideration at some future time, especially if new ideas in sensing come along.

5. How the disambiguator fits in the rest of the system

Some version of a disambiguator program must eventually be included in the system if the robot is ever to deal effectively with complex universes. I envision a hierarchy of control something like the following. In the course of execution some program will decide that it can't live unless it knows for sure what reality is. This may result from a software failure if the program is unable to find the object's exact position in real space or to determine all of its dimensions. This decision may also be based on a hardware signal saying that the space where the hand was to grasp an object was unoccupied, or that the hand ran into an object while moving through a space that was alleged to be unoccupied. In the hardware interrupt situation it need not be true that the object to which the hand objects is the ambiguous one. in fig. 2

if the house-shaped interpretation of the two large objects is selected (AB-CD), the hand will fail to grasp block E because the location of block E was calculated using incorrect information about the surface supporting the stack of small blocks. Since block E is known to be not ambiguous, then to find the ambiguous object which must be resolved it is only necessary in this case to recurse down through the the successive supporting objects of block E until one is found which was originally tagged ambiguous, in this case wedge AB.

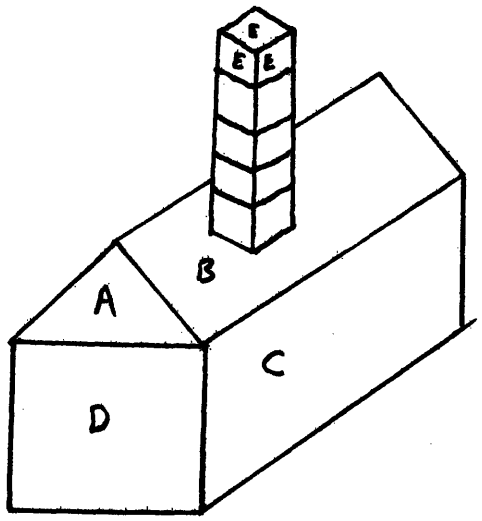

Fig. 2

Once it is decided which object(s) to resolve, control is transferred to the disambiguator program, which will query the body finder about the other plausible alternative and compare the two. To actually resolve the ambiguity, the disambiguator should have at its disposal a repertoire of possible actions to perform. Among these may be included using the hand's sensory devices as an "occupied" predicate, using the hand to pick up an object or a number of objects and then perform a scan, and the type of sliding along the surface motion described above. The disambiguator should have some heuristics to help it decide which of these methods to use. For example, using the hand as an "occupied" predicate is almost always helpful in cases of illusion, but not nearly so often useful when the ambiguity is one of occlusion. This is due to the accessibility of the point to be tested. In cases of illusion there is usually a relatively large volume of space from which to choose a testing point; the point so chosen will usually be easy for the hand to reach; and there are usually no other objects so close to the point that the hand may accidentally knock them down or otherwise become distraught by their presence.

In the present system occlusion and illusion may be distinguished in the following manner. The fact that the program that finds the real-space dimensions and location of an object fails if the object is occluded provides a measure for determining ambiguity by occlusion. In

illusions the same information may be found with no difficulty, but the execution of hand movements to grasp the object will fail if the original interpretation was incorrect.

More generally, to tell whether the ambiguity is illusion or occlusion, the following system-independent heuristic may be desirable. In an illusion there are two (or sometimes more) marrying objects in question in each interpretation, one of which partailly occludes the other. The other objects in the scene, unless they involve another case of ambiguity, are not ambiguous. In occlusion ambiguities it is incidental if the two objects marry. Usually occluded ambiguous objects do not marry other occluded ambiguous objects. If two ambiguous bodies are relatively close to each other, one should check to see if they form a matching-T configuration. This is a very common form of ambiguity by occlusion.

6. Heterarchical considerations

The presence of heterarchy in a complex system of this type is absolutely crucial to its performance. The lack of heterarchy restricts the perceptual-manipulative abilities of a robot to incredibly simple feats. The integration of good heterarchical features in the system can enable it to do a variety of surprisingly intelligent and "human-like" things.

The implementation of some version of the ideas presented here will provide the robot with a new means of interacting with and gaining knowledge from its environment. This is particularly significant to the

manner in which the functions of scene analysis and object manipulation are performed. Scene analysis is done in an outside-in manner. The eye provides information about the environment which is used to construct a model of the universe as the robot thinks it is. The hand is used on an inside-out basis to modify the environment by certain actions. With the exception of a few primitive interrupt signals, the hand is not used at all to get knowledge about the environment. With the implementation of a touch system as proposed here, the hand takes on an interpretive rather than merely a passive role. The scene analyzers have another source of information coming in from the outside. The robot has more interaction with its environment; it obtains a more accurate knowledge of reality; and its intelligence is therefore enhanced. The logical difference between scene analysis and object manipulation is reduced, facilitating a smoother cooperation between the two in the robot's overall performance.

This work is not meant to be definitive. It is merely intended to suggest some general observations which apply to the ambiguity resolution problem. When it comes time to actually implement some programs to resolve ambiguities, another study will be necessary to consider these and other ideas in the light of a different (and hopefully improved) system configuration.

Neither should the reader conclude that vision has found a powerful or even potentially threatening competitor for the job of providing the interface between the real world and its internal computer representation. It is indeed possible that either the vision system or the touch system could be implemented to the exclusion of the other to

perform this function at the level of the present system. But it is only necessary to consider the human system for a moment to realize that these two procedures are complementary, that both working together and communicating with each other are greatly more effective than either by itself. The visual-perceptual and the motor-manipulative systems and the others with which the human system has been endowed were all included in this version of humanity in order to allow the greatest possible flexibility and universality of interaction between the human and his world. This same goal prompted the work reported here.

Appendix: An experiment concerning resolution of ambiguity by humans

The purpose of this experiment was not to determine how people resolve ambiguous scenes in general, for a whole slew of complex abilities are used, making the human system much too complicated to study all at once. Instead, the object was to attempt to characterize how humans might disambiguate scenes if restricted to blind movement of the hand and the sensory devices thereof. There were seven subjects. This number may seem a trifle small, but it was not the intention to characterize the behavior in this situation of all people, but rather to observe the performances of a few people to gain an idea of the variety of ways that people might do this sort of thing.

Each subject was given a short speech about the work on the robot being done here. A problem with which the robot will have to cope, they were told, is the resolution of ambiguous scenes. For various reasons, it was deemed not profitable to use the eye in this process, but rather to be limited only to moving the hand, and to use the hand's sensory devices. They were then told that they would be shown a sequence of ambiguous scenes. For each picture they were to tell what they saw in it, giving alternatives if the scene was ambiguous. Once they determined the ambiguity in each picture, they were asked how they would resolve the ambiguity using only their fingers as touching devices. They could assume that they possessed the ability to move the hand to any desired point in space accurately, even though they were to be blind.

What follows is the sequence of drawings shown to the subjects, after each of which is a description of the subjects' responses.

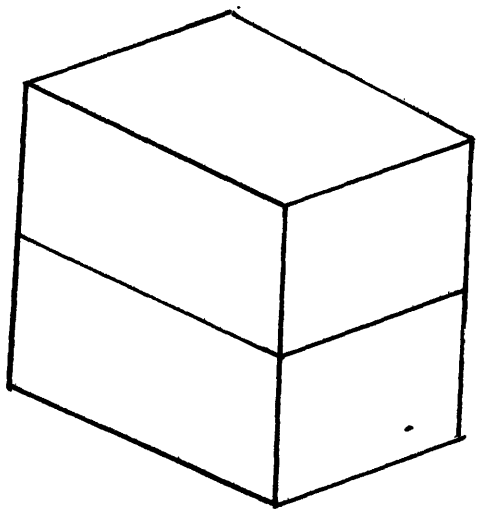

Fig. A-1

The subjects were asked to describe what was present in the scene. Without exception they responded that the picture represented one block resting on top of another block. None suggested that the bottom object could be anything other than a block, although that is indeed possible.

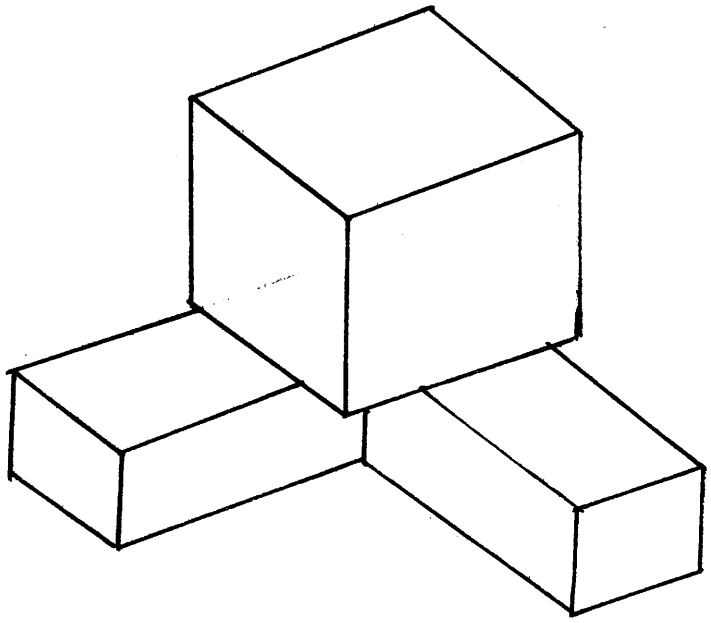

Fig. A-2

The question here is whether there are two blocks sitting directly on the table or whether it is one l-shaped block on which the supported block rests. The responses were:

No. of Responses	Response
2	Follow along the back edge of the object(s) and count whether 1 or 3 corners are present.
2	Test for the presence or absence of the back corner.
2	Move along the back edge of the right hand block from right to left and see how far the surface goes.
1	Follow along the back edge of the object(s) and see if there is a concave edge.

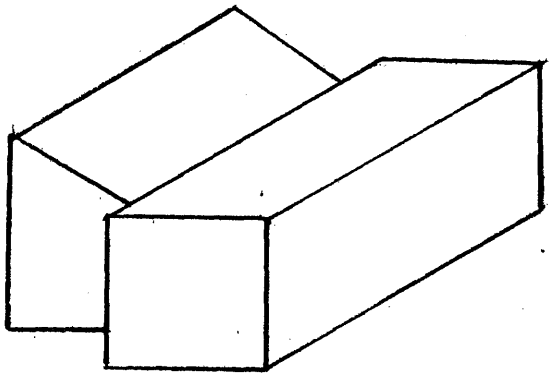

Fig. A-3

Is the object which is occluded by the block a wedge or a trapezoidal solid?

All subjects proposed to run down the inclined surface until a discontinuity in direction was encountered (i.e., until the finger hit the table, the block, or a vertical drop).

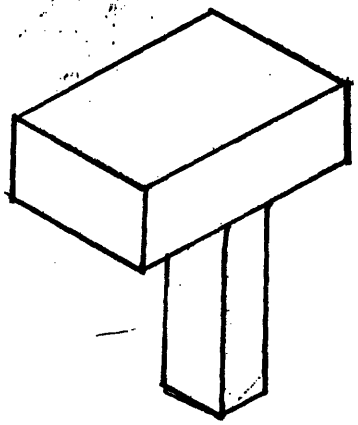

Fig. A-4

What is the height of the standing object?

6 Distance = speed * time , up the side of the standing object.

1 Its height = the distance from the bottom of the top block to the table, if the top block is supported by the standing block.

Does the standing block support the other block?

3 Move vertically along the surface of the standing block and see if you run into something at the top.

3 See if there is a concave or a convex edge at the top of the standing block.

1 Run finger along whole bottom surface of top block.

Is the standing object a wedge or a right rectangular parallelpiped (block)?

5 Count the number of sides around its horizontal perimeter (this is equivalent to counting the

number of changes in direction minus one).

1 Does the sum of the angles of change of direction in a horizontal perimeter sweep equal 180 or 360 degrees?

1 Move finger along back surface and see if there is a change of direction back there.

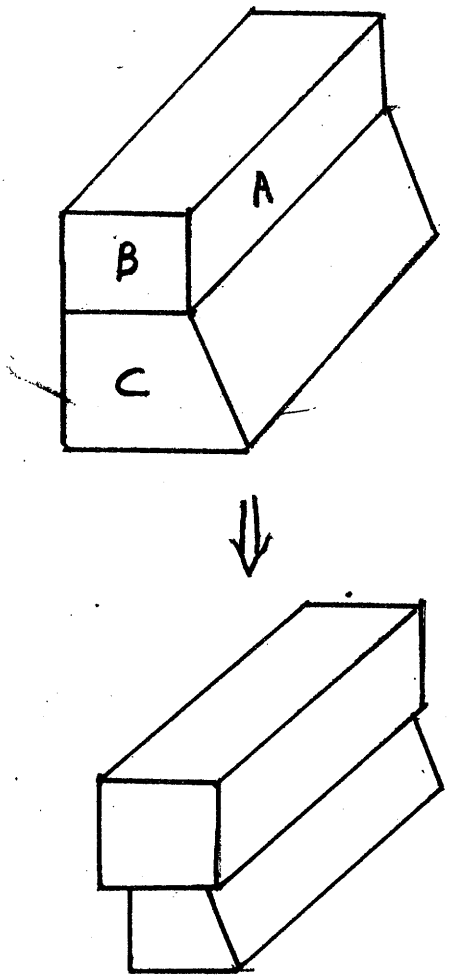

Fig. A-5

Is this a trapezoidal object supporting a block or a block in front of and occluding a wedge or a trapezoidal object?

3	Move finger up inclined surface and measure the change of direction.
2	Move down on surface a.
1	Move down on surface b.
1	Move up on surace c.

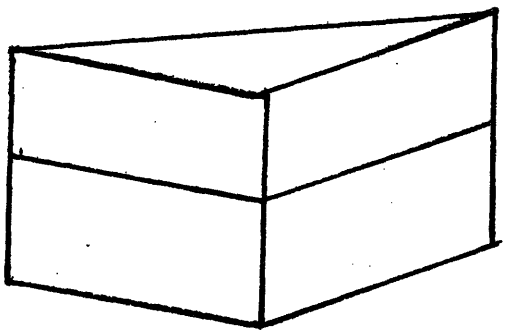

Fig. A-6

Is the bottom object a wedge or a block?

4	Move down the back surface of the top wedge and see if it hits the ground or the block.
3	Count the number of sides of the bottom object.
1	Sum the angles around the perimeter of the bottom object to 180 or 360 degrees.
1	See if the back of the bottom object has one or two surfaces.

(Multiple response from one subject.)

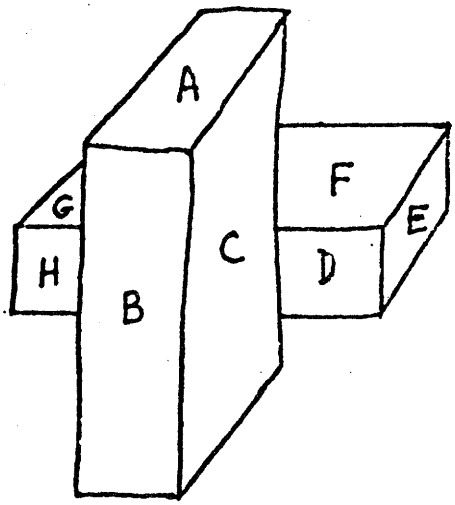

Fig. A-7

Are bodies gh and def distinct or part of the same object?

4 Move along surface g toward f or vice versa and determine whether there is a discontinuity.

3 Move along edge fd toward edge gh and proceed as above.

1 Same as the first method but with surfaces h and d.

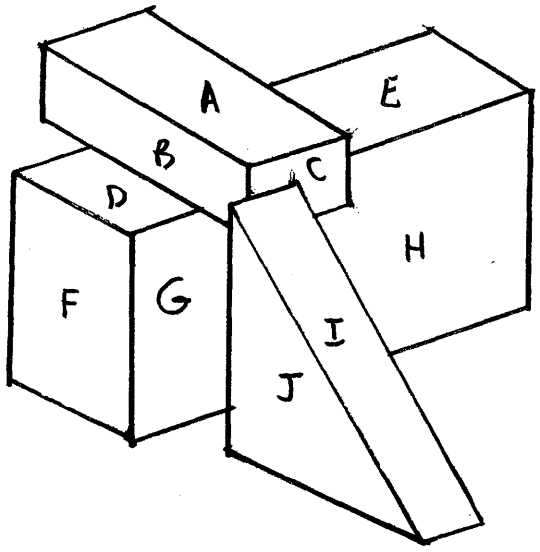

Fig. A-8

are bodies dfg and eh distinct or the same objects?

5 Use the back surfaces as in scene 7.

1 Use the front surfaces as in scene 7.

1 Trace the perimeter around the two (or one) objects.

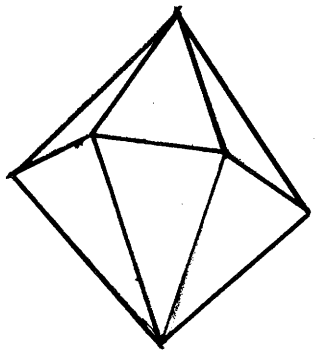

Fig. A-9

The preceding was the sequence of pure occlusion figures shown to the subjects. Before proceeding with the figures of illusion, the subjects were presented with the drawing in fig. 9, and asked how to determine how the regions were connected into bodies. Since this cannot be done merely by touching, the subjects were told that in this case they would be allowed to move the objects and then take another look. The responses of five of them fell into the category of pushing some some object from its place and then looking again to see what came with it. The other two proposed to pull some object from its place and then look again to see what came with it.

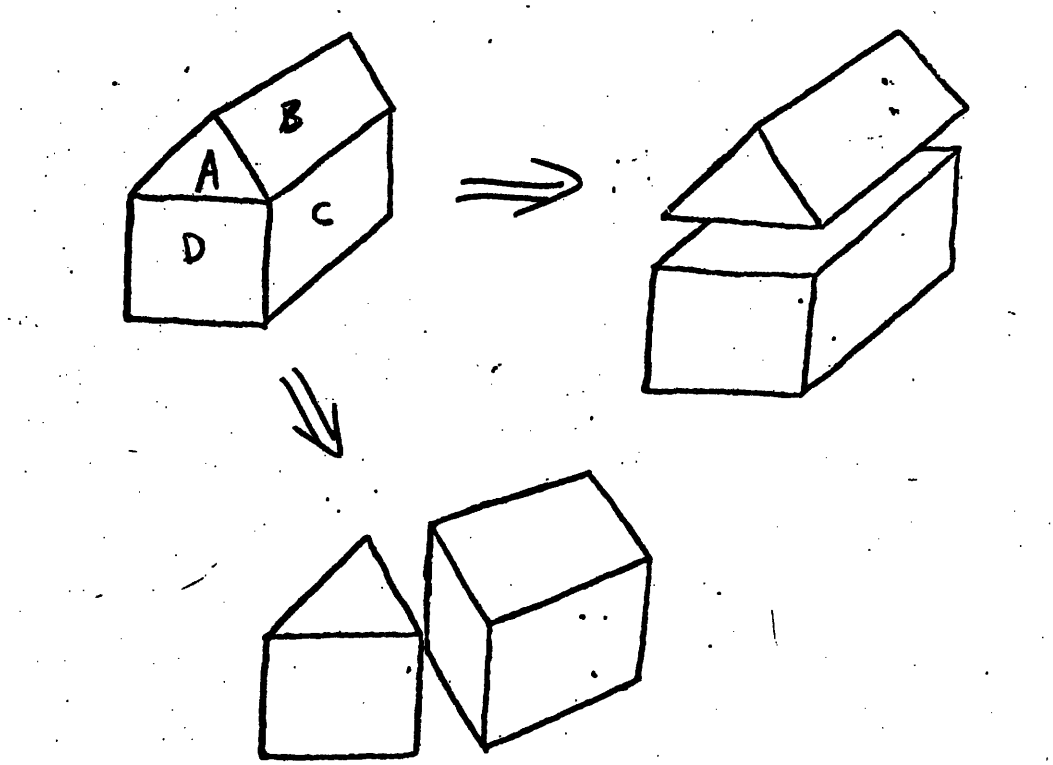

Fig. A-10

BC-AD or AB-CD (i.e., is the scene composed of bodies BC and AD or of bodies AB and CD)?

3	Touch surface B, move from there to surface C and measure the angle between them.
1	Touch C, move to B, continue on, moving to the back sides and eventually to the table top again. Count the number of sides encounterd (3 or 4).
1	Assume that it is a house-shaped object and test for the presence of the edge formed at the top of the two inclined surfaces.
1	Measure the angle between C and D.
1	Measure the angle between a and b.

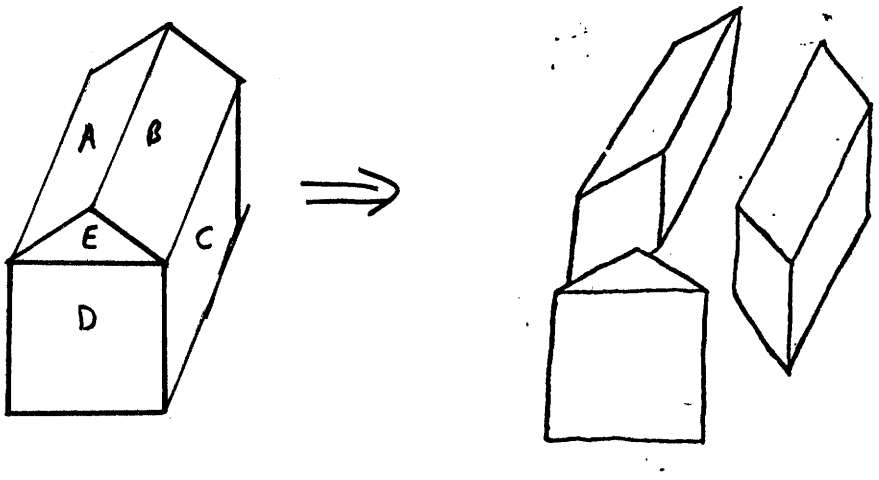

Fig. A-11

ABE-CD or ABC-DE?

3	Measure the angle between surfaces A and B.
2	Trace out the horizontal perimeter to see if it is rectangular or pentagonal.
1	Measure the angle between C and B.
1	Check for the presence or absence of the edge between A and B.
1	Measure the angle between d and e.

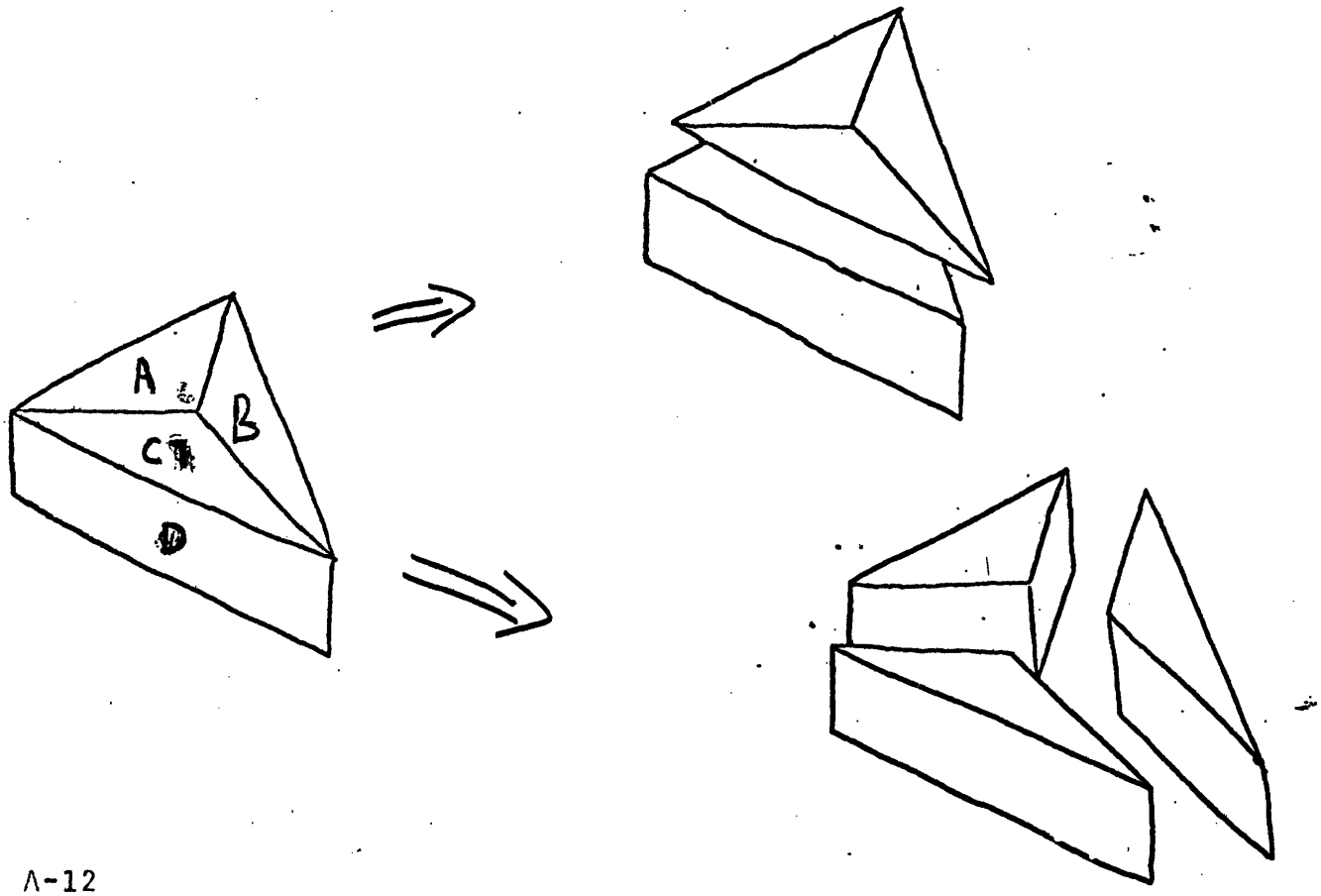

Fig. A-12

ABC-D or A-B-CD?

5	Move around on the surfaces A, B, and C to see if that region is flat.
1	Check for the presence or absence of the raised peak at the point where A, B, and C intersect.
1	Measure the angle between D and C.

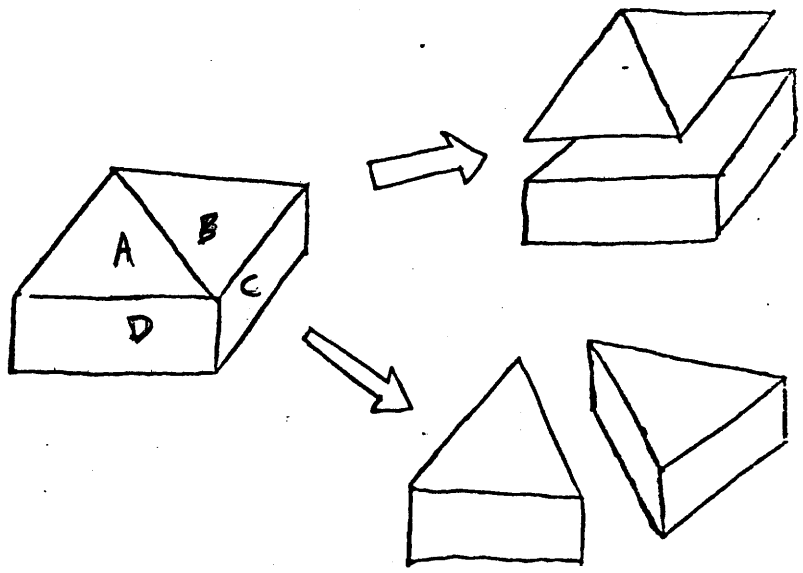

Fig. A-13

AB-CD or AD-BC?

4 Measure the angle between A and B.

2 Measure the angle between D and A.

1 Test for the presence or absence of the peak which would be present in the AB-CD interpretation.

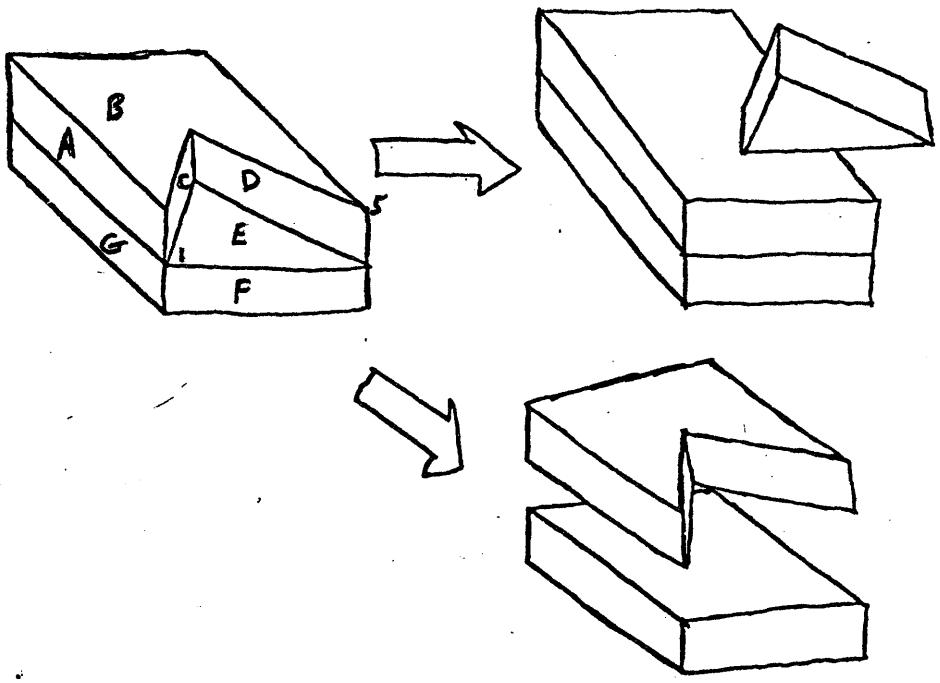

Fig. A-14

AB-CDE-FG or ABCD-EFG?

4	Move from point 1 toward point 2 and see if the space between is empty.
1	Trace the perimeter of the top block and count the number of its sides.
1	Measure the angle between E and F.
1	Try to put the finger in the wedge-shaped hole.

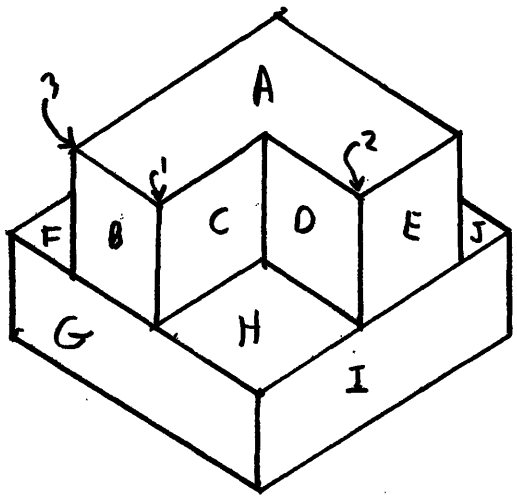

Fig. A-15

ABCDE-FGHIJ or CDH-ABEFGIJ?

2	Move from point 1 toward point 2 and see if the space below is occupied.
2	Try to put the finger in the cube-shaped concavity of the upper block.
1	Move from point 3 toward and then past point 1 testing if the space below is filled or empty when point 1 is passed.
1	Count the number of angles made in making a trace of the horizontal perimeter of the top block.
1	On the A surface, move from the back corner toward the front corner and see if you fall off before the front corner is reached.

In the cases of ambiguity by occlusion the answers given by the subjects seem straightforward and no obvious procedures seem to have been overlooked by the subjects. In the cases of illusion, however, all but one of the subjects refused to use the procedure that seems to be most direct in many of the cases, i.e., to attempt to put the finger in a spot which is occupied in one interpretation but not in the other and to test whether or not there is indeed something there. This indicates that these subjects did not trust the ability to move their hands accurately to a given location in space, even though they were told the could. In spite of this unwillingness to use this method, all used this ability implicitly to move the finger initially to a surface of an object from which they would start they testing motions.

References

1. Ernst, Heinrich A. MH-1, A computer-operated mechanical hand. M.I.T. Ph.D. thesis, Dec. 1961.

2. Rattner, Martin H. Extending Guzman's SEE program. M.I.T. B.S. thesis, July 1970.