VISUALLY INTERPRETING THE MOTION OF OBJECTS IN SPACE

Jon A. Webb[1]
J. K. Aggarwal[1,2]

TR-81-3 June 1981

[1] Department of Computer Sciences
[2] Department of Electrical Engineering
The University of Texas at Austin
Austin, Texas 78712

This research was supported by the Air Force Office of Scientific Research under grant number AFOSR 77-3190.

Abstract

Under certain assumptions, it is possible to recover the structure of a group of points just from several monocular views of them. This process is called structure from motion. Humans have this ability, and it could be used to automatically recover three-dimensional information from the world using a single camera. A new method for recovering structure from motion is developed. The method uses the assumption that each rigid movement consists of a translation and a rotation about a fixed axis. It is also assumed that the points seen are far enough from the camera so that their motion in depth can be ignored.

Visually Interpreting the Motion of Objects in Space

A fundamental ability of the human visual system is its capacity for interpreting motion in space. Investigating this process is justified for psychological reasons alone, but there are many practical benefits to be gained as well. For example, computer applications in robotics and the photogrammetry of human motion would be advanced by even the partial solution of problems relating to visual interpretation of motion. We will concentrate here on the interpretation of the movement of objects in space using a single fixed camera. We will discuss the interpretation of the motion of objects, presenting a new method for interpreting the motion of rigid and jointed objects.

In the discussion following, we will assume that we have the (x,y) image positions of several points over a period of time. We will not discuss how these positions can be obtained automatically; some research in this area is surveyed in [7], and elsewhere in this issue Thompson discusses an algorithm for the recovery of this information. In practice these positions may be obtained interactively from images of objects with identifiable targets. The problem we will consider is how a reasonable three-dimensional structure may be assigned to these points.

It is clear that we cannot determine the three-dimensional structure in absolute terms. This is because anything we see could be produced by a small object close to the camera or a large object farther away. So when we say "determine a three-dimensional structure," we mean assign a three-dimensional structure with an unknown scaling factor which could account for the observed positions of the points. If we wish to find the structure in absolute terms, more information, such as the absolute position of one of the points in the object, will be required. We will also be able to segment points into objects, so that the structure of each of the objects will be determined to within a different scaling factor.

The first question we consider is whether such a structure can be assigned at all, since it is not obvious that the (x,y) positions determine a structure. In fact, they do not in general; some assumptions must be made about how the points move in space. We will discuss these assumptions shortly, but first we will consider the remarkable capacity of human vision for interpreting two-dimensional motion three-dimensionally.

Psychologists have known for some time that humans can interpret a two-dimensional moving figure as if it was seen in three dimensions. The typical psychological experiment showing this displays to the subject a two-dimensional projection (i.e., a movie, a CRT display, or a shadow) of a moving three-dimensional figure. The subject thus obtains only two-dimensional information, whether he uses one or both eyes to watch the figure. Subjects consistently report seeing three-dimensional structure in these conditions, and in the right circumstances they can describe the structure correctly. The first major study of this phenomenon was by Wallach and O'Connell [12], who called it the "kinetic depth effect." We will refer to it as "structure from motion," after Ullman [11].

Structure from motion is commonly used in three-dimensional graphic display systems, which allow the user to input the three-dimensional coordinates of an object, and then examine its three-dimensional structure by moving and rotating the object as it is displayed on a CRT screen. Some of these systems use perspective projection (objects farther away look smaller) or shading to enhance the effect, but this is not absolutely necessary. The motion alone is enough to give the impression of three-dimensionality.

If we could find structure using motion, we would be able to recover three-dimensional information on objects using one camera. At present we must use multiple cameras to get three-dimensional information about objects. Now in order to find structure from motion, some assumptions must be made about the object being viewed, because a sequence of views of a single point moving in space tells us nothing about its position in depth. The position of a point in space is determined by three numbers, while only two are known from its projected position. We want to make an assumption that does not require previous knowledge about the objects we see, since the people in the psychological experiments had no previous knowledge. We can't use shading assumptions, since people could see three-dimensional structure even when isolated points with equal brightness were displayed. The only assumption left is one about the motion of the points.

Mathematically, the motion of any system of points can be written in the form of a transformation that is applied to the points. If the motion is rigid, then the transformation can be decomposed into two parts: a translation and a rotation, which are applied in sequence. Ullman [11] was the first to use assumption of rigidity to interpret the three-dimensional structure of a system of points.

The equations of projection of a system of points relate the unknown three-dimensional coordinates of a point with its known two-dimensional coordinates. If rigidity is assumed, new equations can be written relating the distances between the points to their known two-dimensional projections. Since these distances must be fixed, a new set of equations can be written for each frame, and with enough frames and enough points the equation set will be determined or over-determined, so that the structure can be recovered.

Ullman used the rigidity assumption to recover the structure (to within a scaling factor) of rigid systems of points. In order to simplify the analysis, he assumed the points being viewed were far away, relative to the distance between them and their motion in depth, so that parallel or orthographic projection (distance has no effect on object size) could be used. Ullman was able to find the structure with at least three views of four points, and his solution was in closed form.

Roach and Aggarwal [10] applied the rigidity assumption to images of moving blocks. They did not assume parallel projection, so that their solution was more general. However, this solution was not in closed form; numerical techniques were used to solve the system of equations. Ullman also considered the case where parallel projection could not be assumed.

There are two main problems with this approach. The first is that accuracy in determining three-dimensional structure depends very heavily on accuracy in determining image plane position. It is difficult to estimate the position of a feature point, such as a corner, on the image plane, so that the equations have to be heavily over-determined to find the structure correctly. Also, in both of these studies the feature points could not all lie in a plane (coplanar points lead to a degenerate system of equations) so that finding appropriate points can be difficult in some cases. For example, in athletics images, the points of interest are on the limbs of the subject, so that they are nearly coplanar; in fact, they are almost collinear. The rigidity assumption alone probably is not sufficient to interpret these images.

The second problem is more fundamental. The rigidity assumption alone does not determine the distance between two rigidly connected points. Even worse, we cannot even determine if two points are rigidly connected if we only use the rigidity assumption. This is because the two points can lie anywhere on the two rays that project onto their position in the image plane, as shown in figure 1. Only when there are three or more points can the structure be determined.

It is therefore somewhat surprising to find that people can interpret the motion of just two points in space. Johansson and Jansson [6] showed subjects the two-dimensional image of a moving rod. Subjects saw a rod moving in space. Since the positions of just two points can be recovered here the subjects must be making some other assumption to recover the structure.

Still more surprising are some other experiments by Johansson [4]. In these experiments subjects were shown a movie of a person walking about a dark room with lights attached to his major joints. Only the lights could be seen. There is a very strong impression of three-dimensional motion in these movies. Here again there are only two points on each rigid part of the jointed object being viewed.

In connection with Johansson's experiments, some work in motion vision should be mentioned. Four separate attempts (other than the one to be presented here) have been made to interpret figures like these. The first, by Rashid [9], involved measuring the image plane distance between every pair of points in the image, and then testing for consistency in distance over a large number (25 or 30) of frames. There is some connection between this approach and the one we will present, the most significant difference being that we interpret the image three-dimensionally, while Rashid concentrated on image plane based analysis.

The second approach, by O'Rourke and Badler [8], involved using higher-level knowledge in image interpretation. A detailed model of a human figure was used to interpret the motion of a moving dot pattern like those studied by Johansson. An interesting feature of this research was the the model could be used to interpret structure even when parts of the figure in the image were not visible; for example, the program could predict the position of the figure's hand even when it was out of sight behind the figure's back. We will not be using high-level knowledge in interpreting Johansson's figures, even though humans obviously use such knowledge at some point.

Two other researchers have examined the problems in interpreting images like these. Clocksin [1] developed a heuristic method for recovering the connectedness structure of objects. Hoffman and Flinchbaugh [3] used the assumption of planarity of motion to recover both the connectedness structure and the three-dimensional structure of the objects. Planarity of motion is a special case of the assumption developed here.

We will interpret Johansson's figures as collections of rigid parts, each part consisting of two points (the joints) that are rigidly connected. Our approach is described in greater detail elsewhere [14]. In order to interpret the motion of just two points, an additional assumption must be made. For the reasons given above, the assumption must be about the motion of the points. Under the rigidity assumption, any motion can be written as a transformation in the form A t, where A is a rotation and t is a translation. We will assume that the axis specified by A is fixed over short periods of time. We call this the <u>fixed axis assumption</u>.

Any rigid part whose movements consist of translations and rotations around a fixed axis will satisfy the fixed axis assumption. Because of limitations in design and the presence of gravity, a surprisingly wide variety of natural and man-made objects satisfy it. Rigid objects that travel on the ground must rotate about a fixed axis pointing out from the ground, as do frisbees, tops and maple seeds. A rolling object rotates about an axis perpendicular to its direction of travel and parallel to the surface.

Most jointed object motion satisfies this assumption as well. For example, a ballerina executing a pirouette holds everything but her head and feet still, relative to her torso. Skaters spinning on a skate move their arms, but they only move them towards and away from the axis of rotation, and do so quite slowly relative to their period of rotation. Divers and gymnasts move in quite complicated ways, but generally they rotate about one axis at a time. Normal walking, whether two- or four-legged, moves the limbs in planes parallel to the line of travel. Thus each limb rotates about an axis perpendicular to the line of travel.

We will also make the assumption that Ullman made, i.e., that the points on the objects are far from the camera relative to the distance between them and their motion in depth. This means we are assuming the apparent distance between two points changes only as a result of rotation of one about the other, and not as a result of their combined motion in depth. Interpretation of motion in depth is much harder, because many apparent changes in structure can be due either to motion in depth, or rotation about a point that is fixed in depth, or both. The human visual system apparently makes some further assumptions to analyze motion in depth. These assumptions are not completely understood, and the interpretation of motion in depth by a computer might best be done differently in different contexts.

The fixed axis assumption and the assumption of parallel projection make possible a simple method for determining the structure and motion of any group of rigidly connected points. Consider the motion of two rigidly connected points. Since they stay a fixed distance from each other, the second must stay on a sphere about the first, as shown in figure 2. Under the fixed axis assumption, the second point must stay a fixed distance from an axis passing through the first, so that it lies on a cylinder passing through the center of the sphere, as shown in figure 3. The intersection of the cylinder and the sphere is a circle in a plane normal to the axis.

Under the assumption of parallel projection, this circle stays a fixed distance from the camera, so that the position of one point relative to the other projects onto an ellipse. Were it not for this second assumption, the circle could project onto a spiral as the rigid part got closer to the camera and increased its apparent size. In fact, any curve at all could be produced by rotating the second point about the first and then moving the rigid part towards or away from the camera.

Now the equation describing the ellipse can be discovered by numerically fitting an ellipse to the observed positions of the second point relative to the first. The postion of the second point relative to the first can be shown to be constrained to be on an ellipse in which the minor axis of the ellipse and a line from the origin to the center of the ellipse are collinear. This makes recovery of the ellipse possible from four views of the two points, assuming the complete absence of noise in estimating the positions of the points. In practice, however, more frames are necessary, because there are always errors in estimating the positions of feature points. Once the equation is found, the distance between the two points can be determined because the minor and major axis of the ellipse determine the orientation of the circle that projects onto the ellipse, so that the distance of the center of the ellipse from the origin and the length of the major axis of the ellipse determine the distance between the points. This makes it possible to find the three-dimensional structure of any rigid object that satisfies the motion assumption, even those with as few as two visible points.

This analysis can be extended to the case of jointed objects with two points on each rigid part, such as Johansson's figures. The structure of each rigid part in the jointed object is found first. Then the rigid parts are connected through their joints by solving equations that relate the position of the joint to the position of the rigid parts. This determines the three-dimensional structure of each jointed object.

The same method can be used when there are more than two visible points on the same rigid object. Numerical techniques are used to solve a system of equations that incorporates the fixed axis assumption and the other restrictions. This approach improves the accuracy of three-dimensional interpretation, compared to approaches based on rigidity alone, since more constraints are placed on the system of equations.

In addition to determining structure, we can also segment the points we see into objects, either rigid or jointed. Of course, this segmentation can be done only on the basis of what is seen, so that if two points appear to be rigidly connected the program will interpret them as being connected.

In addition, the structure proposed for the points is not unique; under parallel projection, any structure seen can be reflected in depth and produce the same image. Other information is necessary to determine the correct three-dimensional structure from the two possible structures found by this method. Some of this information could come from the image, such as knowledge about when rigid parts should hide each other from view; other information may involve world knowledge, such as knowing which ways a person's legs can bend.

The structure from motion method developed in this paper has been implemented and run on data similar to the figures studied by Johansson [4]. Full three-dimensional information was available for this data, so that the accuracy of the results could be tested.

Figures 4 and 5 show two sets of data that have been analyzed by the program. The first figure shows a walking man, while a woman swinging a baseball bat is shown in the second figure. Six points are shown on the walking man: the shoulder, the elbow, the wrist, the hip, the knee, and the ankle. These points are shown over a period of 0.26 second. These data were collected using a stereo infrared camera system designed by Eric Antonson of the Mechanical Engineering department at MIT.

The baseball data shows 12 points: a point on the bat, the wrist, the left and right illiac crests, the hips, the knees, the ankles, and the metapharsal joints in the feet. The data shown, extending over a period of 0.32 second, were collected using stereo visible light cameras by Steve Messier of the Department of Health and Physical Education at the University of Texas at Austin.

The parallel projections from the viewpoint shown were given to a program that used the fixed axis assumption to interpret the three-dimensional structure of the figure. The program determined which points moved as if they were connected and estimated the three-dimensional distance between rigidly connected points. The resulting connectedness structures shown in figures 6 and 7 illustrate the constrained nature of human movement when viewed over short periods of time. The connectedness structure for the walking man is correct, except that the elbow is connected to the hip. The connectedness structure for the baseball player has many wrong connections. The program has difficulty determining the connectedness structure of heavily constrained motion because the constraints make the actual movements involved more rigid, so that the connectedness structure has many spurious connections. This is complicated by the fact that the data is noisy, so that the small deviations from rigid motion cannot be used to determine that the motion is not rigid. A program which looked at the sequences over longer periods of time and intersected the proposed connectedness structures determined over short periods of time would be able to determine connectedness structures more reliably.

Tables I and II show the three-dimensional rigid part lengths as determined by the program compared with the three-dimensional rigid part lengths determined by the stereo camera systems. The walking man data is analyzed correctly, but many errors are present in the baseball data analysis. These errors arise for several reasons: (1) the baseball data is more noisy; (2) some of the motion in the data, such as the bat motion, does not satisfy the fixed axis assumption; (3) points that are very close together, such as those in the hip -- illiac crest structure, are heavily affected by small errors in position; (4) points that do not move, such as those on the left foot, cannot be analyzed correctly. However, it appears that this method is reliable enough to aid in the interpretation of some athletic movement, especially when only monocular images are available.

We have developed a system for the interpretation of many moving figures that humans can also interpret. Can we learn psychologically interesting things from this system? Suppose that the human visual system uses the fixed axis assumption to interpret motion. This might help to explain how people can interpret the motion of many objects; we have seen that this method may be powerful enough to interpret figures like those studied by Johansson. In fact, Johansson himself [5] has studied the role of simple rotations in the interpretation of visual motion. What about motion that does not satisfy the fixed axis assumption? We cannot be sure that there are not other processes that could help people interpret this kind of motion. However, we would not expect motion not satisfying this assumption to be difficult to interpret if the human visual system used rigidity alone to interpret visual motion. The evidence on this point is not clear; the only study we are aware of that studied motion that was not fixed axis is that of Green [2], who showed subjects collections of rigidly moving points. In some circumstances, the points looked less rigid when they moved with a tumbling motion. This effect was present even when there were several points visible. Further studies should be done comparing the interpretation of fixed axis motion with that which is not fixed axis.

For some degenerate kinds of motion the fixed axis assumption is not powerful enough to guide the recovery of structure from motion. This occurs when the axis of rotation either points directly at the viewer (so that the ellipse which is traced degenerates to a circle) or the axis is perpendicular to the line of sight (so that the ellipse degenerates to a line segment). In the second case, the motion can still be interpreted three-dimensionally if it is assumed that the maximal apparent distance between the points occurs when the line from the first point to the second is perpendicular to the line of sight, as suggested by Webb [13]. This assumption has been observed by Johansson and Jansson [6], who showed subjects the image of a rotating rod.

In summary, we have considered the problem of interpreting the motion of points in space from a two-dimensional view of these points. We briefly surveyed the ways of interpreting motion three-dimensionally, and then concentrated on interpreting the motion of objects. We have presented a new technique for interpreting three-dimensional structure. This technique makes it possible to interpret scenes that could not be interpreted previously without high-level knowledge, such as Johansson's figures. We applied this technique to images similar to those studied by Johansson, and found that it was powerful enough to recover structure. Green's experiments suggest that humans may use this or a similar assumption. Two important problems not addressed here are the interpretation of motion in depth and the low-level vision problem of the extraction and correspondence of the positions of feature points.

Acknowledgments. The authors acknowledge the helpful comments of Larry Davis and Worthy Martin, both of the Department of Computer Science. The authors would like to thank Steve Messier, of the UT Department of Health and Physical Education, who provided the baseball data, and Don Hoffman, of the MIT AI Lab, who provided the walking man data. We would also like to express our appreciation for the comments of an anonymous reviewer and Wes Snyder, the guest editor of this issue, which improved the presentation in this paper.

REFERENCES

[1] Clocksin, W F, "Inference of structural descriptions from visual examples of motion: preliminary results," DAI Working Paper 21, Dept. of Artificial Intelligence, University of Edinburgh, May, 1977.

[2] Green, B F, Jr., "Figure coherence in the kinetic depth effect," *J. Exp. Psych.*, *63*, (3), 1961. pp. 272-282.

[3] Hoffman, D D and B E Flinchbaugh, "The interpretation of biological motion," MIT AI Memo 608, December, 1980.

[4] Johansson, G, "Visual motion perception," *Scientific American*, November 1976, pp. 76-88.

[5] Johansson, G, "Visual perception of rotary motions as transformations of conic sections," *Psychologia*, *17*, 1974. pp. 226-237.

[6] Johansson, G and G Jansson, "Perceived rotary motion from changes in a straight line," *Perception and Psychophysics*, *4*, (3), 1968. pp. 165-170.

[7] Martin, W N and J K Aggarwal, "Survey: Dynamic scene analysis," *Computer Graphics and Image Processing*, *7*, (1978). pp. 356-374.

[8] O'Rourke, J and N I Badler, "Model-based image analysis of human motion using constraint propagation," *IEEE PAMI*, *PAMI-2*, (6), November 1980. pp. 522-536.

[9] Rashid, R F, "Towards a system for the interpretation of moving light displays," IEEE PAMI, PAMI-2, (6), 1980. pp. 574-581.

[10] Roach, J W and J K Aggarwal, "Determining the movement of objects from a sequence of images," IEEE PAMI, PAMI-2, (6), 1980. pp. 554-562.

[11] Ullman, S, The interpretation of visual motion, The MIT Press, Cambridge, MA, 1978.

[12] Wallach, H and D N O'Connell, "The kinetic depth effect," J. Exp. Psych., 45, (4), 1953. pp. 205-217.

[13] Webb. J A, "Static analysis of moving jointed objects," Proc. First National Conference on Artificial Intelligence, Stanford, California, August, 1980. pp. 35-37.

[14] Webb, J A and J K Aggarwal, "Visual interpretation of the motion of rigid and jointed objects," to be published.

From	To	Est. Length	Act. Length	Rel. Error
		(Metres)		
Shoulder	Elbow	0.344	0.335	2.53%
Elbow	Wrist	0.283	0.274	3.35%
Shoulder	Hip	0.584	0.579	0.808%
Hip	Knee	0.438	0.437	0.175%
Knee	Ankle	0.435	0.437	1.90%

TABLE I. Rigid part lengths, walking man data

From	To	Est. Length	Act. Length	Rel. Error
		(Inches)		
Bat	Wrist	23.7	27.3	13.3%
Right ill. crest	Left ill. crest	926	8.55	***
Right ill. crest	Left hip	***	10.3	***
Right ill. crest	Right hip	2.64	2.23	18.2%
Left ill. crest	Left hip	11.7	2.34	399%
Left ill. crest	Right hip	3190	9.03	***
Left hip	Right hip	***	10.5	***
Left hip	Left knee	16.6	16.9	2.07%
Left knee	Left ankle	14.6	15.5	5.69%
Left ankle	Left meta. joint	22.8	3.97	473%
Right hip	Right knee	17.5	17.1	2.81%
Right knee	Right ankle	14.8	15.3	3.79%
Right ankle	Right meta. joint	3.72	4.21	11.6%

TABLE II. Rigid part lengths, woman with baseball bat. Asterisks indicate very large errors.

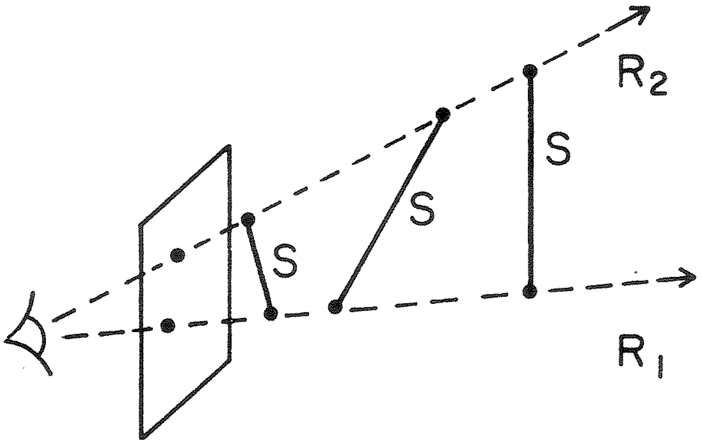

Figure 1. S can lie anywhere between the two rays.

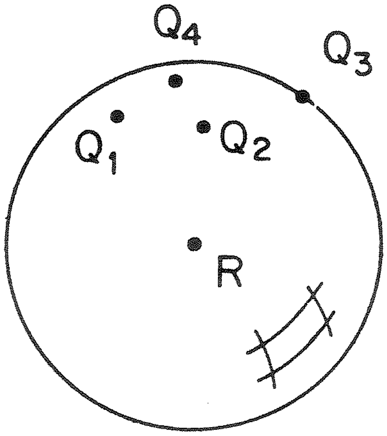

Figure 2. The second point must lie on a sphere about the first

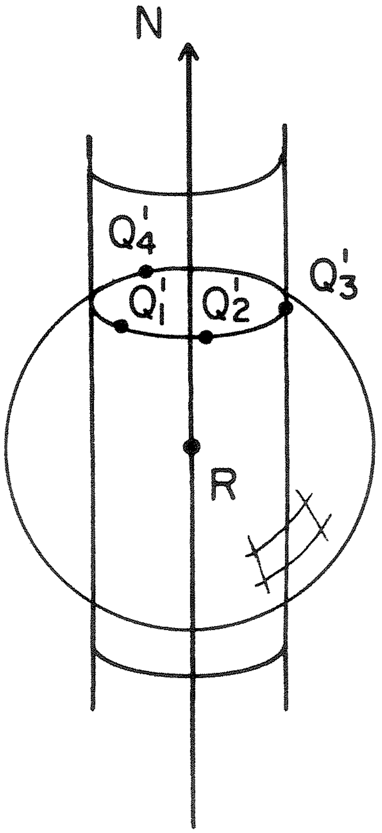

Figure 3. The second point must stay a fixed distance from the axis

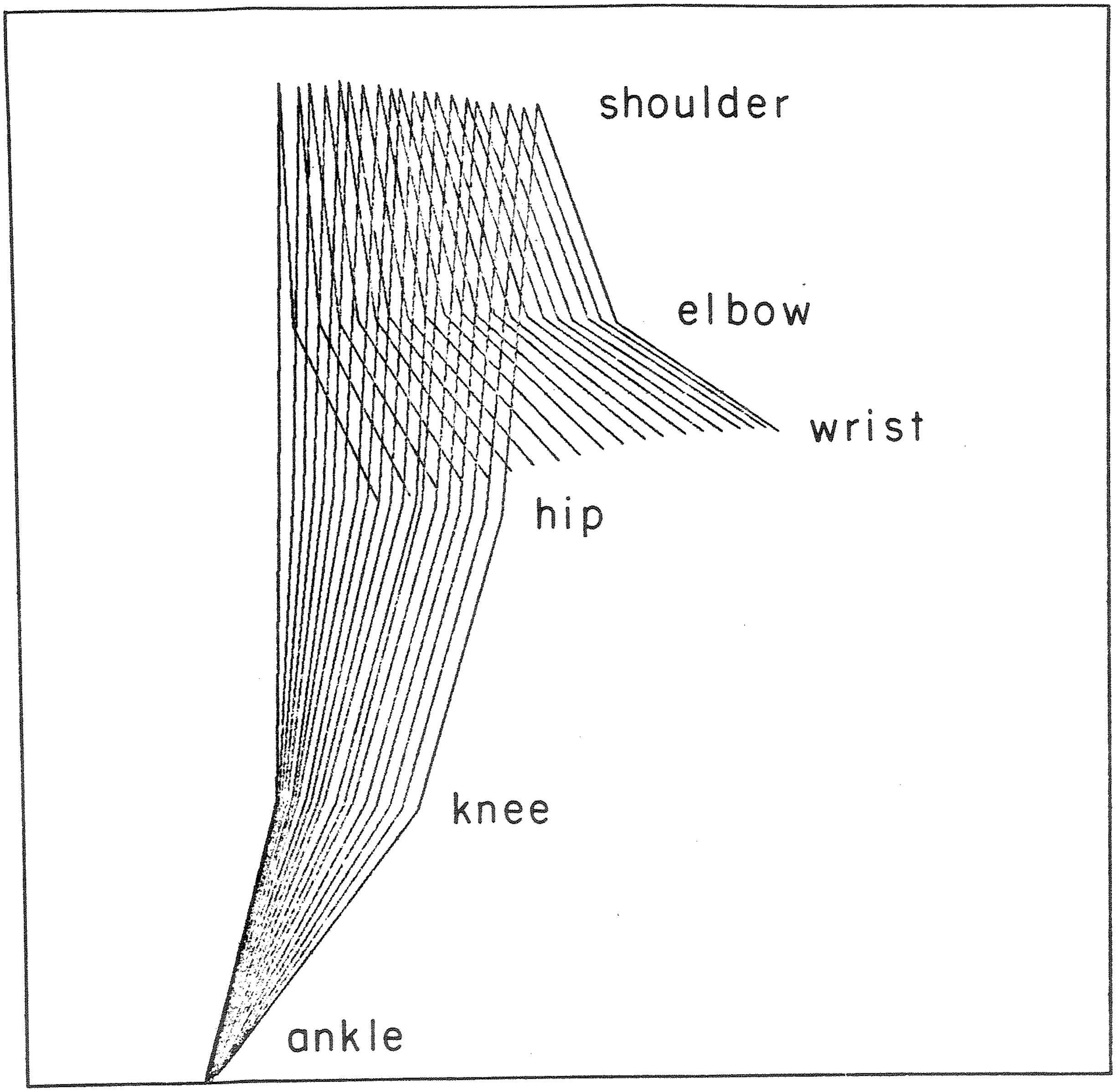

Figure 4. The walking man

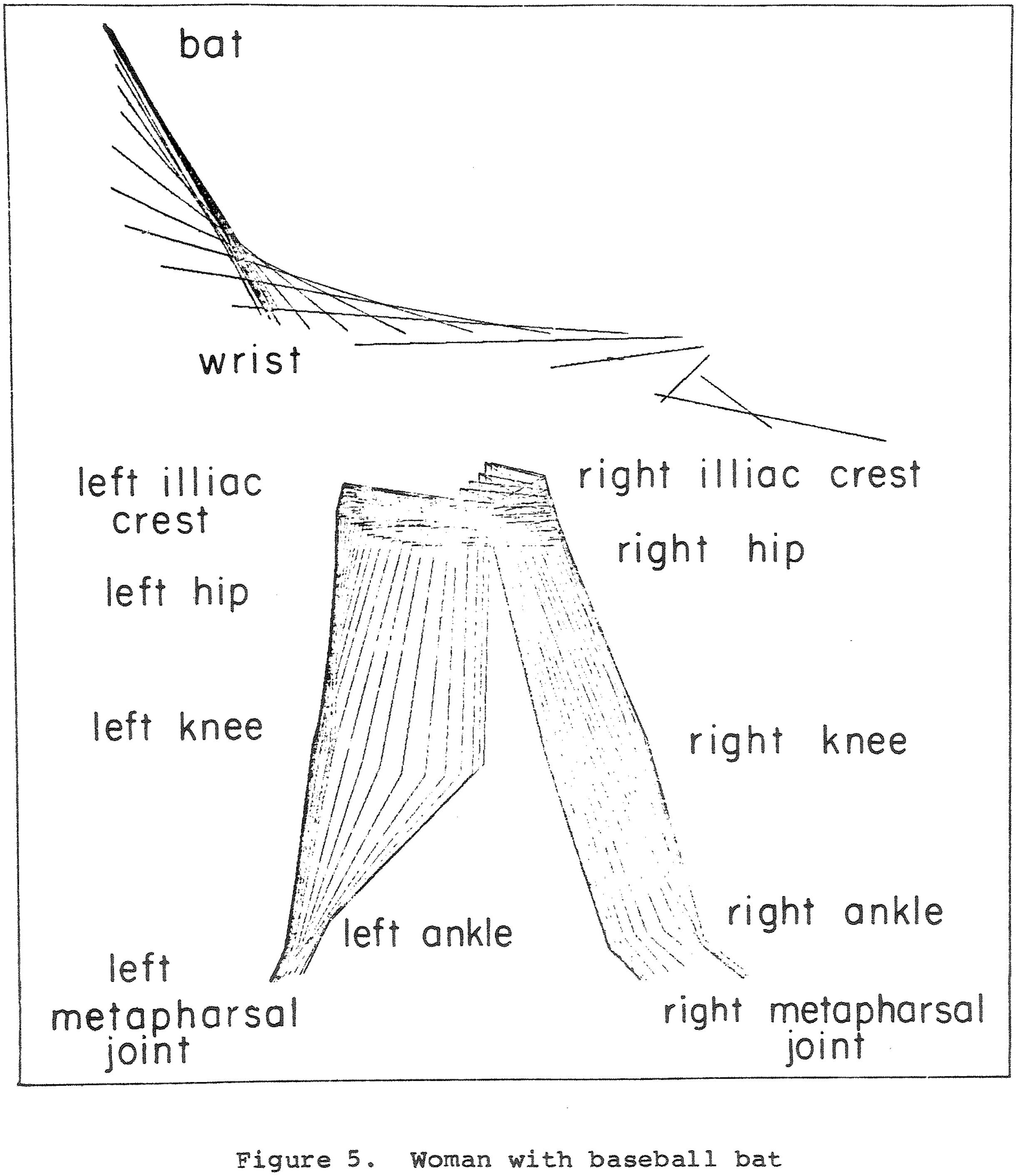

Figure 5. Woman with baseball bat

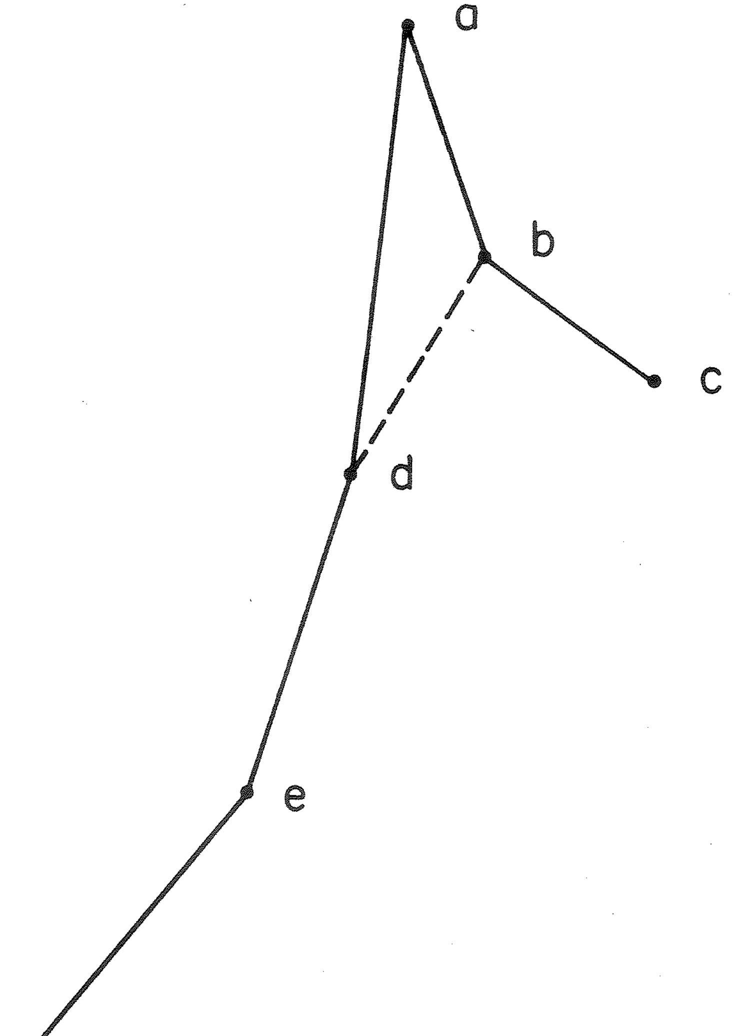

Figure 6. Connectedness structure for walking man: a -- shoulder; b -- elbow; c -- wrist; d -- hip; e -- knee; f -- ankle. Dashed line shows incorrect connection proposed by program.

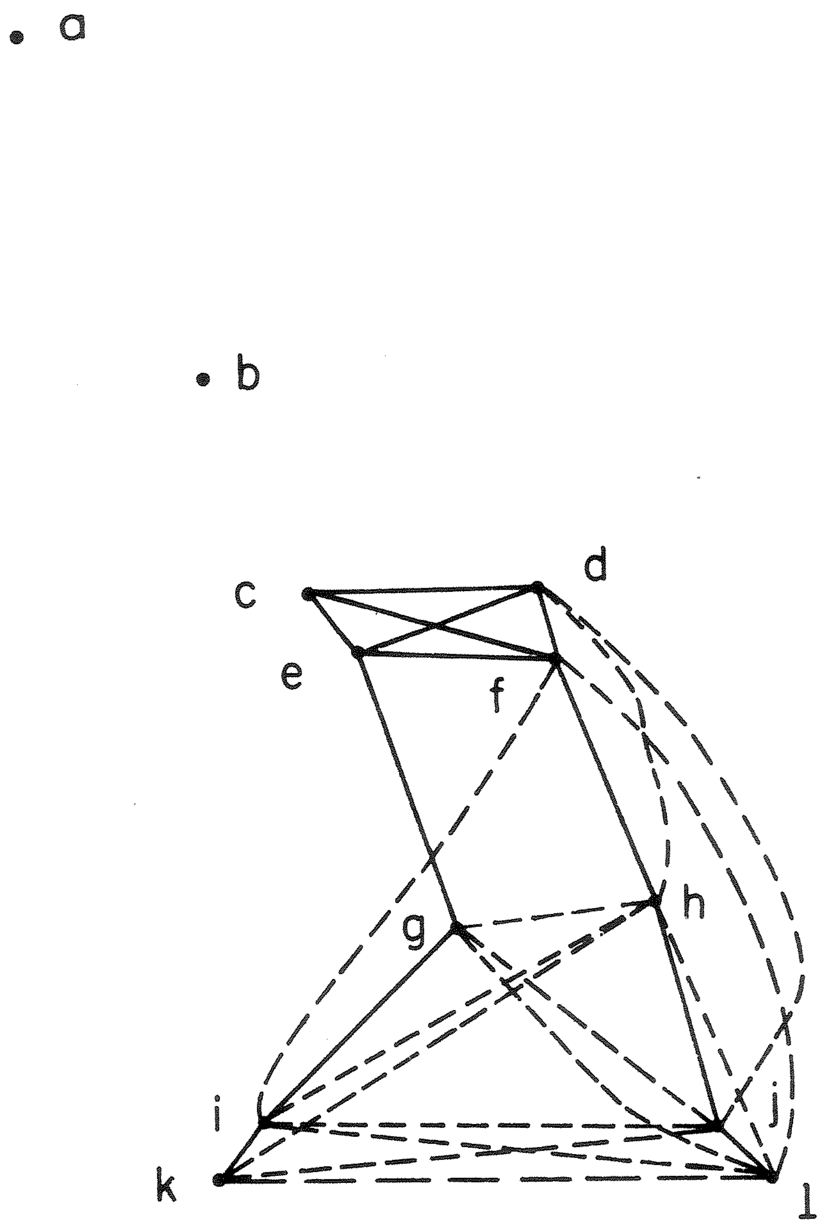

Figure 7. Connectedness structure for woman with baseball bat: a -- bat; b -- wrist; c -- left illiac crest; d -- right illiac crest; e -- left hip; f -- right hip; g -- left knee; h -- right knee; i -- left ankle; j -- right ankle; k -- left metapharsal joint; l -- right metapharsal joint. Dashed lines show incorrect connections proposed by program.

Errata

Page 12, line 11, "metapharsal" should be "metatarsal"

Page 17, Reference 4 should read:

[4] Johansson, G., "Visual motion perception," Scientific American 232 (6), June 1975. 76-78.

Figures 5 and 7 throughout - "metapharsal" should read "metatarsal".