COMPUTER INTERPRETATION OF HUMAN STICK FIGURES

Martin Herman
Department of Computer Science
Carnegie-Mellon University
Pittsburgh, PA 15213

ABSTRACT

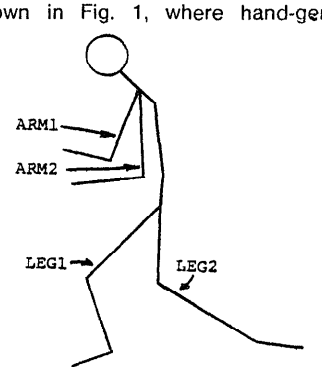
A computer program which generates context-sensitive descriptions of human stick figures is described. Three categories of knowledge important for the task are discussed: (1) the 3-D description of the figures, (2) the conceptual description of the scene, and (3) heuristic rules used to generate the above two descriptions. The program's representation for these descriptions is also discussed.

1. Introduction

This paper describes a computer program, called SKELETUN, which generates context-sensitive descriptions of 2-D, static, human stick figures. The motivating interest is to study the process of extracting information communicated by body postures. Stick figures have been chosen to approximate the human form because they eliminate the problems involved in processing fleshed-out human figures (e.g., extracting them from the image, identifying and labeling body parts), yet they maintain the overall form conveyed by gross body posture.

SKELETUN currently operates in two domains, emotions (figures may be sad, happy, depressed, etc.) and baseball (batting, catching, etc.) Its knowledge of baseball is much more complete, however. It can accept any figure and interpret it in terms of the following baseball activities: (1) batting, (2) throwing, (3) running, (4) catching a high ball with one or both arms, (5) catching a ball at torso height with one or both arms, (6) fielding a grounder with one or both arms.

An example of how a figure is interpreted in the baseball domain is shown in Fig. 1, where hand-generated English



1. Two arms catching torso-high ball (very good confidence)
2. Batting (fair confidence)
3. Two arms fielding grounder (poor confidence)

Fig. 1a

PHYSICAL DESCRIPTION

The figure is in a vertical orientation with the feet below the head.	((vertical orientation) verygood) ((feet to bottom of head) verygood)
The figure is facing left and the face is pointing left. The torso is bent forward. The	((facing left) good) ((face pointing left) good) ((torso bent forward) good)
elbow of arm1 is in-middle and down. It can be considered either as partly or half bent. The	((elbow1 is in-middle) good) ((elbow1 is down) verygood) ((elbow1 partly bent) good) ((elbow1 half bent) good)
elbow of arm2 is in-middle and down. It can be considered either as partly or half bent.	((elbow2 is in-middle) good) ((elbow2 is down) verygood) ((elbow2 partly bent) good) ((elbow2 half bent) good)
The knee of leg1 is forward and partly bent.	((knee1 is forward) verygood) ((knee1 partly bent) good)
The knee of leg2 is down and partly bent.	((knee2 is down) verygood) ((knee2 partly bent) good)
The lower body is in a configuration similar to "feet well planted."	((both legs "feet well planted" cfig) good)
The figure can also be considered in a diagonal orientation with the feet to the lower right of the head (but with lower confidence than vertical).	((diagonal orientation) good) ((feet to lowerright of head) good)
In this case, it is facing lower left with the face pointing lower left. The following then changes from the previous description: the elbow of arm1 can be considered as either down or forward. The knee of leg2 is forward.	((facing lowerleft) fair) ((face pointing lowerleft) fair) ((elbow1 is down) good) ((elbow1 is forward) good) ((knee2 is forward) verygood)

MEANING-BASED DESCRIPTION

The figure is catching a ball at torso height with two arms, with very good confidence. It may	((two-arms-catching-torso high-ball) verygood)
also be viewed as batting, but with only fair confidence.	((batting) fair)
Finally, it may be fielding a grounder with two arms, but with only poor confidence.	((two-arms-fielding-grounder) poor)

Fig. 1b

descriptions are shown alongside the computer-generated output. SKELETUN's primary purpose is to generate a description of what is communicated by body posture - the "meaning-based" description. In the process of generating this description, it also provides the 3-D configuration of the figures - the physical description. Briefly, the notation in the example is as follows. If a figure is viewed from the front or back, each elbow or knee can be either out from the torso, in to the torso (i.e., crossing the torso) or in-middle (i.e., along the same line as the torso). If the figure is viewed from the side, each elbow or knee can be either up, forward, backward, or back-up (i.e., backward and up). All assertions in the descriptions have discrete confidence values.

The input to SKELETUN is a hand-encoding of the x, y coordinates of the end points of the line segments of each figure, plus the center of the circle representing the head. SKELETUN assumes that all figures are complete and valid, and that no objects other than figures are in the scene (a scene may have two figures).

This paper gives an overview of the types of information conveyed by gross body postures, SKELETUN's representation for this information, and some inference rules used to generate this information from 2-D scenes. See [7] for details.

1.1 Background

This work views vision as a medium of communication, recognizing that an important goal of the visual process is to provide the viewer with a "meaning" description of the external world.

Most scene analysis systems are primarily concerned with identifying objects and other entities in a scene and specifying the spatial configuration of these entities [6, 3, 8, 4]. Given a scene with human figures, such systems would tend to identify the individual figures, their body parts, and other objects, and then specify the spatial relationships of these entities [9, 10, 1]. SKELETUN goes one step further in the interpretation process. It tries to determine what the people are doing, and perhaps why they are doing it.

Although some previous work has taken the point of view of vision as communication [2, 14, 15], their primary purpose was to analyze and describe motion scenes, rather than to study how body posture conveys information.

2. Knowledge categories

Five categories of knowledge have been identified as important in the process of generating descriptions of 2-D scenes of stick figures. The first three represent important levels at which the scene should be described.

- Two-Dimensional Description - a low-level description involving the direction of each body part (each is a straight line segment), the angle of each joint (in the 2-D plane), and body parts which overlap (required for establishing touching relationships).

- Physical Space Description - a 3-D description of the physical configurations of the figures.

- Meaning Space Description - a description in terms of the information communicated by the figures (e.g., running, fighting, crying). The concepts here are said to be in Meaning Space since "meaning" (or "conceptual" information) is extracted from the scene.

The next two categories involve knowledge used to extract the physical and meaning space descriptions from the 2-D description.

- Human Physical Structure - information dealing with the various parts of the stick figure body and components of these parts.

- Inference Rules - heuristic rules used to obtain the 3-D configuration of the figures from the 2-D scene, and to determine what the figures are doing based on the 2-D and 3-D configurations of the limbs.

The following sections will further discuss the 2nd, 3rd, and 5th categories. More details than can be provided here on all of the categories may be found in [7].

3. Physical Space Description

In order to infer what is being communicated by a figure's body posture, there must be knowledge of at least part of its 3-D configuration, for it is a 2-D figure interpreted as being in 3-D space to which meaning is applied.

It is convenient to have two different levels of physical space descriptions. One, called the lower level physical space description, deals with the 3-D positions of the individual body parts. The second, called the higher level physical space description, deals with frequently occurring positions of groups of body parts. Only the first description will be discussed in this paper (see [7] for more details).

Although a figure's 3-D configuration may be represented many ways, the representation to be described next was chosen for two reasons:

1. Its purpose is to describe the figure in a manner useful for generating meaning-based interpretations. If the resolution is too fine (as in [10]), it will contain much information not significant for the task, thus burdening the system. If the resolution is too coarse, it will not contain enough information to perform the task.

2. It is convenient for SKELETUN to be able to express a figure's 3-D configuration in a manner easily understood by humans. The current representation makes this kind of information explicit.

3.1 Descriptions relative to the torso

The 3-D descriptions in SKELETUN are object-centered, as opposed to viewer-centered. That is, locations and directions of parts of the figure are indicated with respect to the figure, rather than the viewer. A viewer-centered description depends not only on the figure being described, but also on its orientation. An object-centered description, however, depends only on the figure being described, resulting in a smaller set of possible descriptions [11].

Accordingly, the positions of the upper arms and legs are represented relative to the torso, and the shape of the torso is represented relative to the overall orientation of the figure. SKELETUN uses the predicates OUT, IN-MIDDLE, and IN to describe the position of each elbow or knee as viewed from the

front, and UP, FORWARD, DOWN, BACKWARD, and BACK-UP to describe the positions as viewed from the side. These predicates are adequate to completely specify (within the resolution of the representation) the 3-D position of any elbow or knee (i.e., upper arm or leg). SKELETUN uses the predicates BENT-FORWARD and BENT-BACKWARD to specify how the torso joints are bent.

3.2 Hierarchy of object-centered descriptions

The positions of the lower arms and legs are represented relative to the upper arms and legs, respectively. Note that a representation of the lower limbs relative to the torso would result in a much larger set of possible descriptions than a representation relative to the upper limbs, since a different description of the lower limb would be required for each position of the upper limb relative to the torso, even if the position of the lower relative to the upper limb were to remain constant.

Since similar arguments apply to describing positions of other body parts, such as hands, fingers, feet, etc., we conclude that each body part should be represented relative to the part it is connected to, resulting in a hierarchy of descriptions [10].

SKELETUN represents the positions of the lower arms and legs by specifying the 3-D angle of the elbow and knee joints. The predicates used are PARTLY-BENT, HALF-BENT, FULLY-BENT, and NOT- BENT.

3.3 Orientation relative to viewer

Thus far, all descriptions have been relative to parts of the figure. The whole figure must also be placed in 3-D space, relative to the viewer. The predicate ORIENTATION describes the overall orientation of the figure either as vertical, horizontal, or diagonal. Given one of these orientations, the predicate DIR-OF-FEET-TO-HEAD specifies the direction of the feet relative to the head. Finally, the predicates DIR-FACING and DIR-FACE-IS-POINTING specify the direction the figure is facing and the direction the face is pointing.

3.4 Physical space inference rules

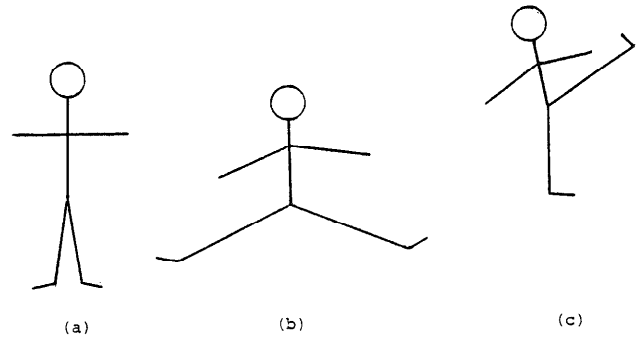
These rules generate the physical space description. They are domain-independent, for they depend only on the 3-D configuration of the figures. As an example of the knowledge in these rules, consider how SKELETUN determines the overall orientation of the figure. A figure is horizontal if both feet are east or west of the head (as in lying). A figure is diagonal if both feet are southeast, southwest, northeast, or northwest of the head.

There are two types of vertical orientations, upright and upside-down. (SKELETUN currently cannot handle upside-down figures.) Fig. 2 shows three extremes of upright figures. In Fig. 2a, both feet are south of the head. In Fig. 2b, both feet are not south of the head; the point midway between the feet is south of the head. In Fig. 2c, the midway point is not south of the head; only one foot is south of the head. Rules which determine whether a figure is upright must examine these three types of cases. For more details on these and other inference rules, see [7].

4. Meaning space description

4.1 Representation

Meaning space concepts in SKELETUN are not represented explicitly in terms of simpler concepts and relationships between them (as in Conceptual Dependency [13]), since SKELETUN's



Three upright stick figures.
Fig. 2

concern is not to extract all the details of each concept. Instead, they are represented as labels (RUNNING, CRYING,WALKING, etc.), where the meaning is represented implicitly in terms of the inference rules which may assert the concept and those which may use the concept to assert other concepts. This is because SKELETUN's concern is to discover and make use of relationships among concepts [12].

Two important classes of information that can be extracted from the body postures of stick figures deal with (1) the physical states of the figures (running, walking, throwing, standing, etc.) and (2) the mental or emotional states of the figures (weeping, happy, thinking, etc.).

Two types of physical states can be distinguished, active and passive. Active physical states involve activities requiring motion, such as running, dancing, or hitting. Passive physical states involve no motion; examples are standing, pointing, and watching.

Mental-emotional states can also be divided into two categories, negative and positive. Negative states generally involve feelings or tendencies such as painful excitement, destruction, dullness, loneliness, discomfort, tension, incompetence, dissatisfaction, and helplessness (e.g., anger, sadness, apathy, panic, hate, grief, disgust). Positive states generally involve feelings or tendencies such as vitality, empathy toward others, comfort, and self-confidence (e.g., cheerfulness, enjoyment, happiness, hope, love, pride) [5]. The negative and positive states can each be further subdivided into passive and active. These will not be pursued here (see [7]).

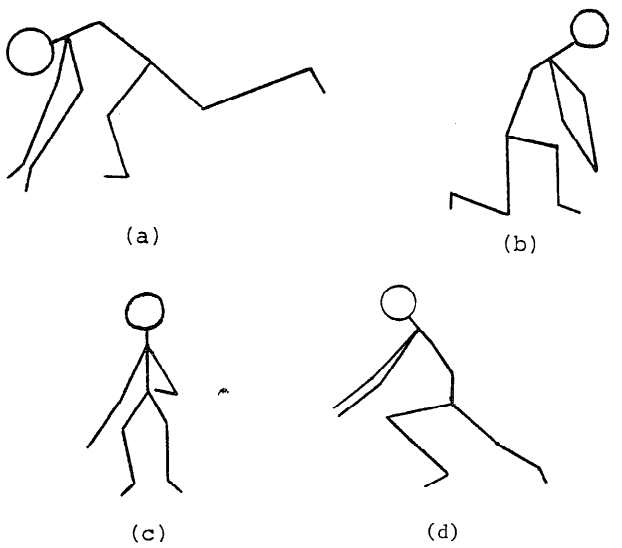
4.2 Meaning space inference rules

These rules generate the meaning space description. They tend to be domain-dependent, since most meaning-space concepts are applicable only in limited domains. As an example of the knowledge in these rules, consider how SKELETUN determines that a figure is fielding a grounder (assuming that the domain is baseball). (See Fig. 3 for examples.) First, one or both arms must be in a "fielding grounder" configuration (a higher level physical configuration described in [7]). In addition, the lower body should be in a configuration similar to "kneeling on one knee" (Fig. 3b), "kneeling on both knees", "feet well planted" (Fig. 3c), or "crouching" (Fig. 3d) [7] and the figure should be vertical. If the figure's orientation is diagonal, its lower body should be in a "crouching" configuration and it must be facing either lower-left or lower-right. Finally, if both arms are in a

"fielding grounder" configuration and the figure is running, it is also fielding a grounder, i.e., running after a ground ball (Fig. 3a).

Acknowledgement

This research is part of the author's Ph.D. thesis done at the University of Maryland, under the guidance of Chuck Rieger and Azriel Rosenfeld. The support of the National Science Foundation under Grant MCS-76-23763 is gratefully acknowledged, as is Mike Shneier for valuable comments, and Ernie Harris for help in preparing this paper.



Each figure is fielding a grounder.
Fig. 3

References

1. Adler, M. Computer interpretation of Peanuts cartoons. *Proc. 5IJCAI*, Cambridge, MA, 1977.

2. Badler, N. I. Temporal scene analysis: conceptual descriptions of object movements. Tech. Rept. 80, Dept. of Computer Science, University of Toronto, 1975.

3. Bajcsy, R., and Joshi, A. K. A partially ordered world model and natural outdoor scenes. In *Computer Vision Systems*, Hanson and Riseman, Ed.,Academic Press, 1978.

4. Barrow, H. G., and Tenenbaum, J. M. MSYS: A system for reasoning about scenes. Artificial Intelligence Center Technical Note 121, Stanford Research Institute, 1976.

5. Davitz, J. R. *The Language of Emotion.* Academic Press, 1969.

6. Hanson, A. R., and Riseman, E. M. VISIONS: a computer system for interpreting scenes. In *Computer Vision Systems*, Hanson and Riseman, Ed.,Academic Press, 1978.

7. Herman, M. Understanding body postures of human stick figures. Tech. Rept. 836, Computer Science Center, University of Maryland, College Park, MD, 1979.

8. Levine, M. D. A knowledge-based computer vision system. In *Computer Vision Systems*, Hanson and Riseman, Ed.,Academic Press, 1978.

9. Marr, D., and Nishihara, H. K. Spatial disposition of axes in a generalized cylinder representation of objects that do not encompass the viewer. AIM 341, MIT, 1975.

10. Marr, D., and Nishihara, H. K. Representation and recognition of the spatial organization of three-dimensional shapes. AIM 416, MIT, 1977.

11. Nishihara, H. K. Intensity, visible surface, and volumetric representations. *Workshop on the Representation of Three-Dimensional Objects*, Univ. of Pennsylvania, Philadelphia, PA, 1979.

12. Rieger, C. Five aspects of a full-scale story comprehension model. In *Associative Networks: The Representation and Use of Knowledge in Computers*, N. Findler, Ed.,Academic Press, 1978.

13. Schank, R. C. Identification of conceptualizations underlying natural language. In *Computer Models of Thought and Language*, Schank and Colby, Ed.,W. H. Freeman and Co., 1973.

14. Tsuji, S., Morizono, A., and Kuroda, S. Understanding a simple cartoon film by a computer vision system. *Proc. 5IJCAI*, Cambridge, MA, 1977.

15. Weir, S. The perception of motion: actions, motives, and feelings. Progress in Perception Research Report No. 13, Dept. of Artificial Intelligence, University of Edinburgh, 1975.