MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

Working paper 118 January 1976

Knowledge Driven Recognition of the Human Body

Glen Speckert

This paper shows how a good internal model of the subject viewed aids in the visual recognition and following of key parts. The role of knowledge driven top-down tools and methods is shown by recognizing a series of human figures drawn from Eadward Muybridge's collection of 1887. Knowledge of the subject's structure and actions are used to find the head, shoulder, elbow, hip, knees, and ankles of the subject.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-75C-0643-0005.

The primary goal of this project is to demonstrate how much a good internal model of the subject viewed aids a computer vision program in visual recognition, tracking, and analysis of key parts. A secondary goal is to learn about how a person moves when he/she is walking normally. Thus the domain is limited to pictures of persons walking. We want to view an ordinary series of pictures that were not made especially to be viewed by a computer program, a series void of easy to locate markers or other artificial clues in an effort to see how much information can be extracted using only knowledge about the subject. Other more precise methods exist for studying motion, for example, multiple flash strobe pictures easily record on one frame the past positions of key points, and is perhaps a better way to analyse motion, but again, the primary goal is not to alanyse the motion, rather to learn if a computer program can actually identify body parts. Since it would be difficult to produce a series of photographs without introducing any bias that would aid in viewing, a well known photographic series produced in 1887 by Eadweard Muybridge was chosen instead.

Muybridge is perhaps the father of motion pictures, and is well known as the photographer who settled a bet between Leland Stanford and George Clark as to whether a horse ever leaves all four feet off the ground at the same time while trotting, by taking a series of pictures of a horse trotting some of which show the horse completely suspended in the air. Muybridge published many sequences of various animals in motion as well as many sequences of humans doing everything from "walking normally" to

"falling prone and aiming a rifle". He set up a battery of from 7 to 24 cameras in a row, and was able to develop a mechanical triggering system which triggered each sequentially with a constant time difference between pictures. The result is similar to a series of consecutive frames of a motion picture, except that the imaginary motion camera can be thought of as moving along the line of physical still cameras, and each frame is taken from a separately aimed and focused camera. These pictures are not nearly as consistent from frame to frame as one would get with a modern movie camera, thus providing a good domain for a program which attempts to find and follow key points.

The series chosen is labled "Man Walking at Ordinary Speed," and appears as plate 3 in The Human Figure in Motion by Eadweard Muybridge, a subset of his works which was published in 1955. It shows a nude male walking normally, viewed from the subject's right, with seven pictures spanning slightly more than half a stride. This series was chosen due to the fact that it is the best plate for viewing with regards to contrast, focus, and consistency. The seven pictures are shown in Figure 1. The performance of the program on one frame from plate 4 is also shown for generality.

Overall Organization

The program assumes that the picture is a five sector by two sector picture of a right-side-up person moving forward (i.e. to the right of the

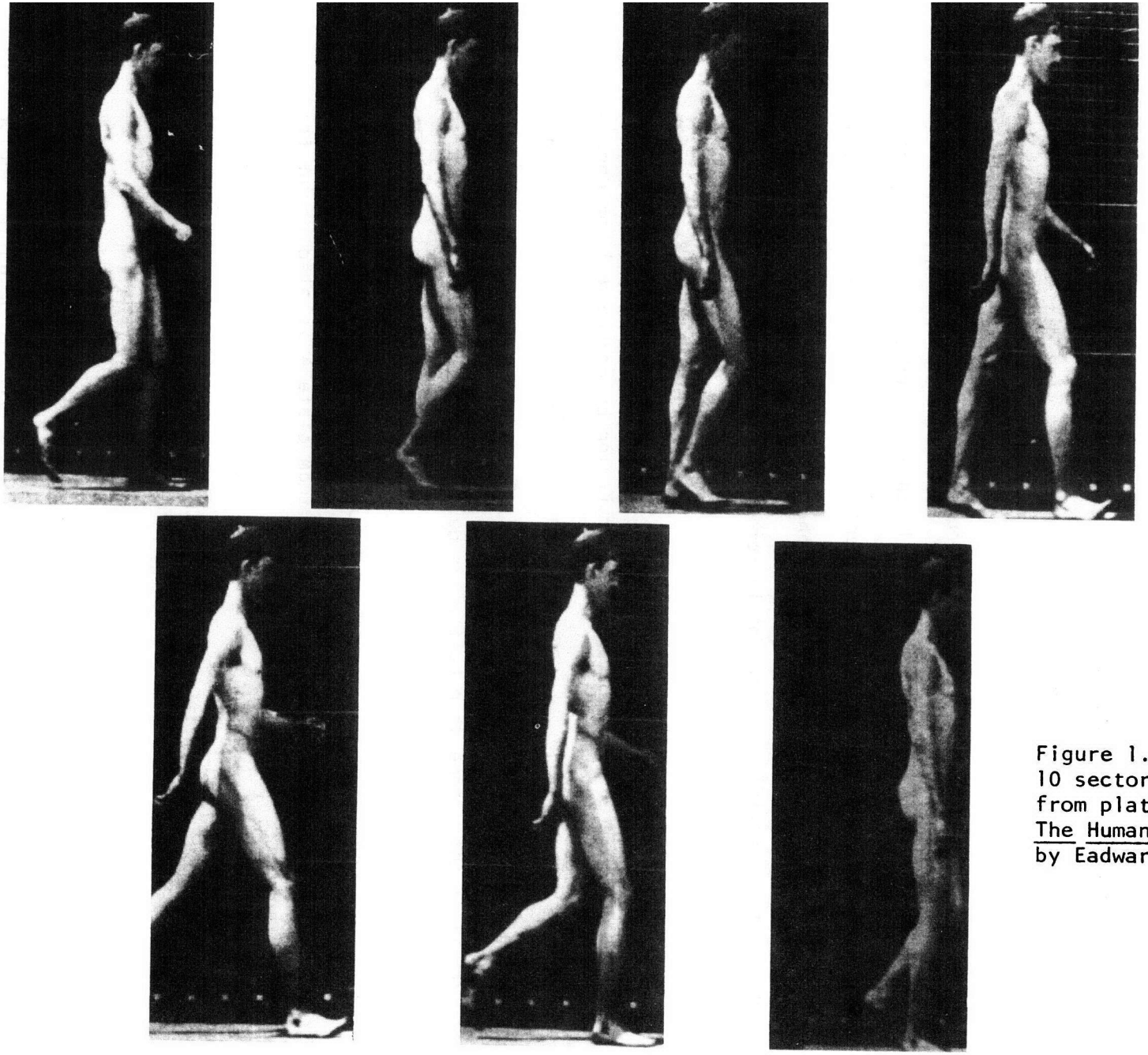


Figure 1. Digitized 10 sector pictures from plate 3 of The Human Figure in Motion by Eadward Muybridge

picture). (If this is not thought to be general, a prescanner could be written which determines if the person is right side up, or upside down, walking right or left, and it could flip the pictures to the correct orientation.) The program consists of several "specialists", such as a head specialist, a shoulder specialist, each of which finds (or attempts to find) one body part based on information gathered thus far, and on its model of how the part it is looking for fits in with the rest of the picture. For example, the elbow specialist knows that the elbow is rigidly attached to the shoulder, and thus need only look in a very constrained area defined as an arc of points which are approximately the right distance from the shoulder. Trying to find the elbow without first finding the shoulder is a lot more difficult, even given the proper model. Each specialist (head, shoulder, elbow, wrist, hip, knees, feet), adds any information gained to a common pool for use by other specialists. If desired, a second pass could be made which would utilize the information obtained in the first pass, thus making available much more information to the early specialists. This was not done due to the resonableness of results obtained in a single pass.

Structural Model

The machine's model of the structure of the subjects is of primary importance, just as for a human viewing the picture. A person knows that the knee is connected to the ankle in spite of other intervening parts (the

other leg, for example). A person can probably pick out most key points by looking at any single picture, knowing the structure of the subject. But occasionally if it is difficult to tell whether the front foot is the right or the left foot (as in a silhouette, or low contrast picture), then more information is needed. If one knows that the subject is walking normally, and that his right arm is swung forward, it follows that the left leg will be the forward one. This type of knowledge about walking is also used in this system. If the frames were closer together in time, then the position in the last picture, along with a calculated velocity and acceleration from previous pictures, could also be used in locating key points, but with seven pictures, this is not much of an aid.

Basic Tools

In searching for the body components, the original picture is viewed directly for the most part, and not, say a Laplacian transform of the original as was done with good success in Japan by Kanade ("Picture Processing System by Computer Complex and Recognition of Human Faces", Kyoto University, November 1973) while viewing faces. Many transformations were tried, including Laplacian, horizontal and vertical differencing, etc., but they seemed to do little except eat computation time and confuse the issue. The most effective tools seem to be a very simple line finder, a simple region grower that first buckets the picture into 8 levels, then grows all regions simultaneously (in a manner similar to John Conway's game

of Life), and a directable ridge follower. The fact that these tools are used in a top-down manner is the key to their success. The program knows what it expects to see, and is merely using the tools for verification, modification, and precision, as opposed to a program which uses its tools for exploration and recognition of the subject.

The line finder is told where to look and how strong a line to look for. The knowledge of where the lines should be is given, and the line finder merely verifies its presence and locates it precisely. Because this is a top-down, or knowledge driven system (the program knows it is looking for a line that is part of a knee) rather than a bottom-up or line finder driven system (line finder finds lines and program identifies them as part of a knee), a simple line finder is sufficient. The line finder merely looks for a change of intensity greater than the given threshold over a distance of four cells (the average sharpness of a line is 3-4 cells). The line finder can also be used to detect how smooth or level a given line is.

The region grower first buckets all the data into eight buckets (eight is arbitrary), and then looks to see how many neighbors are in the same bucket. Depending on the number of neighbors (eight possible) in its bucket, a cell, Z, will either remain the same, or change to the bucket that the plurality of its neighbors are in. Thus if Z has neighbors X,Z,X,Y,Y,Y,W,Y, it will change to a Y. This has the advantage that it changes the whole picture simultaneously, and thus doesn't favor any region over any other.

The ridge follower simply follows a ridge in a downward (decreasing Y) direction by looking at the three cells below it and going

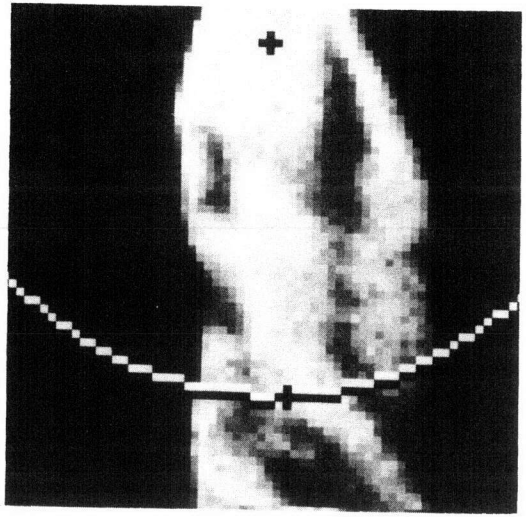
to whichever is highest. The calling program determines which way to go on ties, whether or not to include a fourth point to the right (or left) of the three, and whether or not a horizontal movement is allowed if this far right (left) point is the highest. Thus the calling routine can essentially say, "Follow the ridge, but expect it to go to the right," or "steep right", etc., thus making the ridge follower also a top directed tool.

The brighter (whiter) a point is, the "higher" it can be thought of in a three dimensional plot with intensity as the altitude on an X Y plane. Thus a peak is a local white point, and a well is a darker than average point.

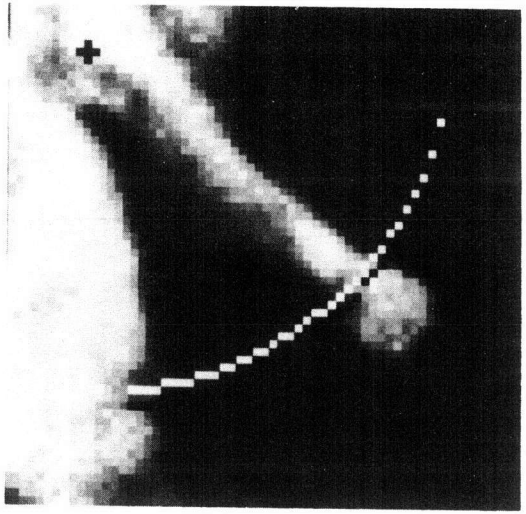
Using this analogy, the line finder looks for cliffs, both up and down, and the region grower cuts a landscape into its contour map, then operates on the map to maximize plain sizes and minimize number of steps, while attempting to maintain the basic information inherent in the original map. The ridge follower steps along a local high ridge for a gven distance. The program views pictures only of size 64 x 64, or one sector at a time. This is due to the fact that this size is the standard sector size for all picture oriented system software on the Micro-Automation system.

These tools are all used in a top-down manner guided by as much knowledge about the structure and actions of the subject as possible. The ridge follower can be seen in action in Figure 2, pictures D and E. The line finder searches along the arcs shown in A, B, and F. C shows the roundedness of the subject's rear end being used to find the hip.

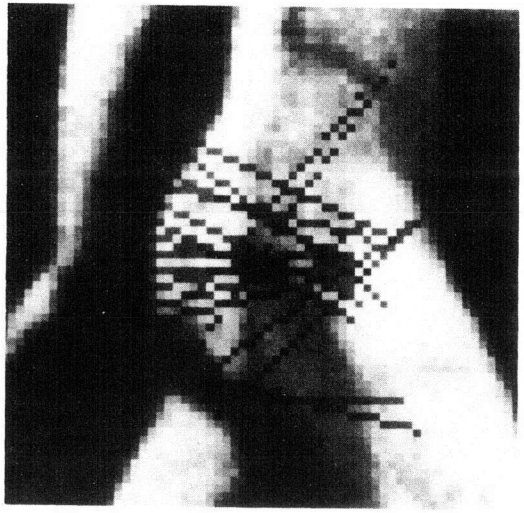
Figure 2



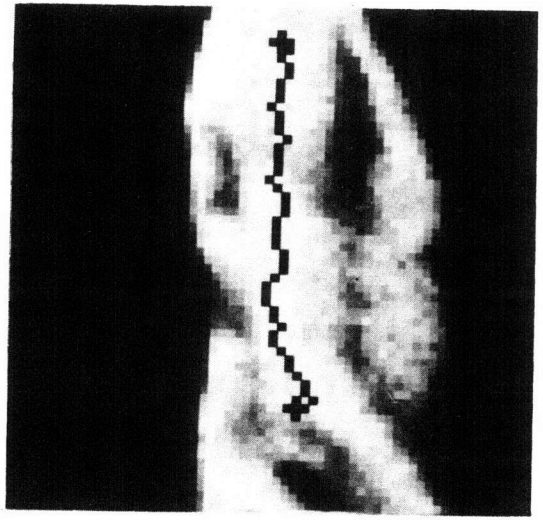
A



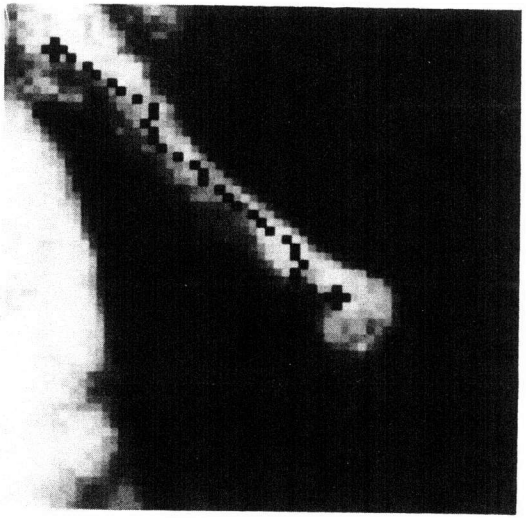
B



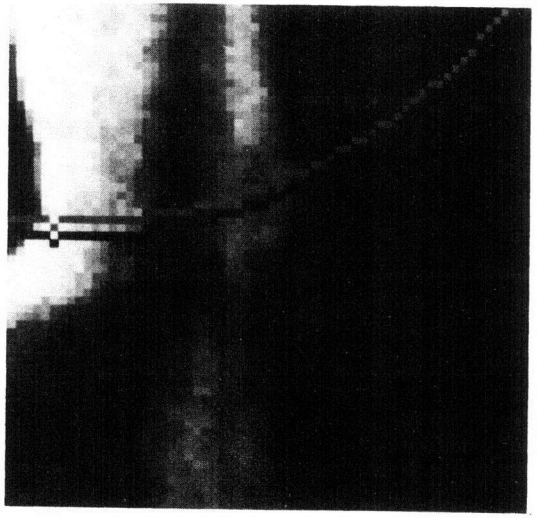
C



D



E



F

Head Specialist

The head specialist has in some sense the most difficult task, since there is only a priori knowledge available at this time, It proceeds in two steps. The first step is to view a compressed, lower resolution picture (a one sector picture of the entire body), and get a rough idea of where the head is, and copy a one sector close up picture about its best guess from the disk into core. Most of the specialists behave in a similar manner, since one sector can be viewed by the user through a crt display, and since all parts to be found are less than one sector in size. The first part of the head specialist uses the region growing algorithm described above, and then essentially thresholds the picture into the body of the person and the darker background. The head is then looked for near the top of this roughly thresholded "body". (If one objects to the assumption that the person must be lighter than his background, one can easily add a prescanner which decides whether the background or the subject is lighter.)

Once the basic body has been found, the head is assumed to be at the top, and a one sector picture "close up" is viewed. The chin is an important feature, as it determines the 'head-height', a valuable measuring length used in finding other parts. The line of the front of the neck (assumed facing to the right) is followed up until it snakes back upon itself (see figure 3a). If it doesn't snake back upon itself, as in figure 3b, the point of least slope is used. The chin is then verified by a

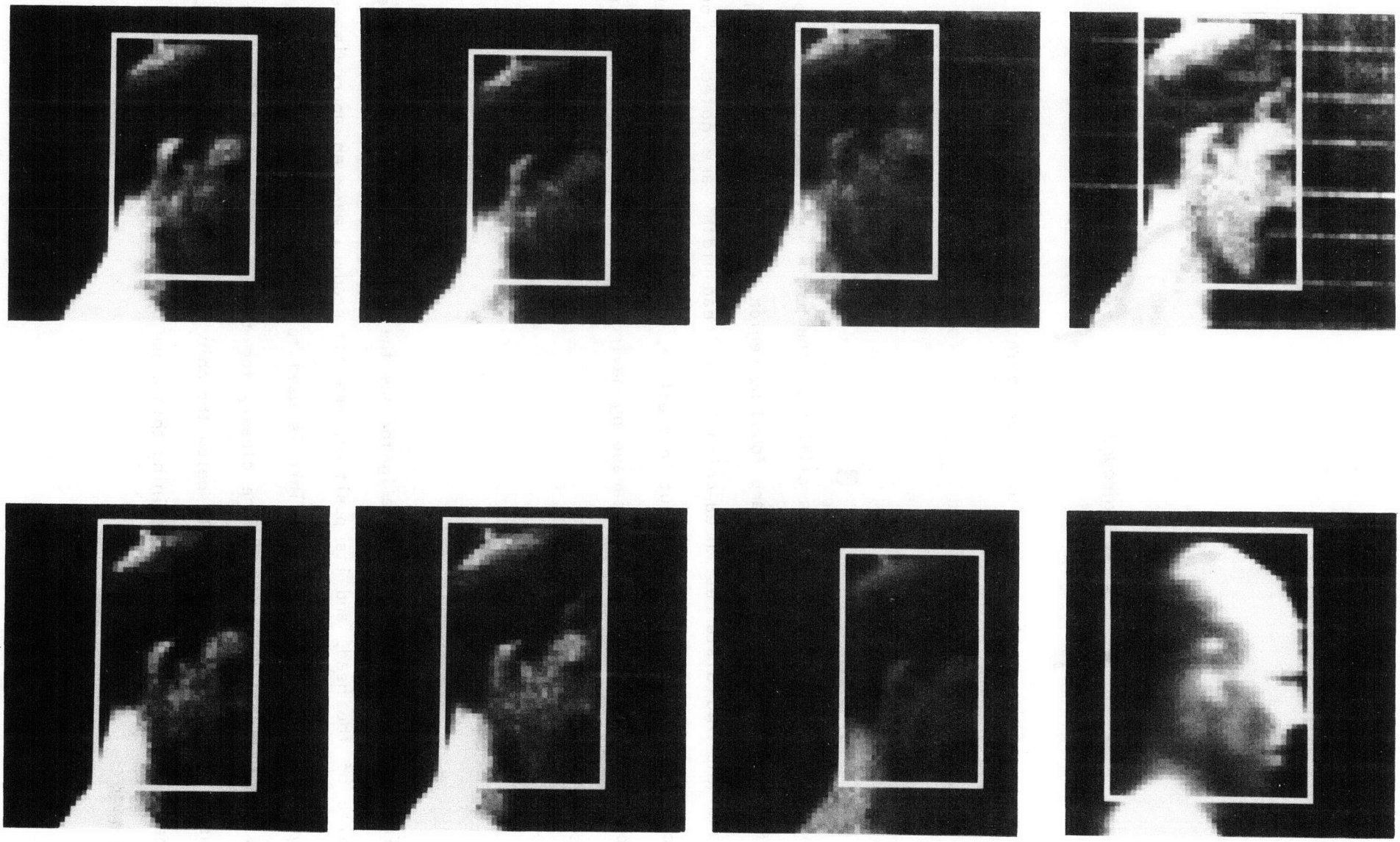
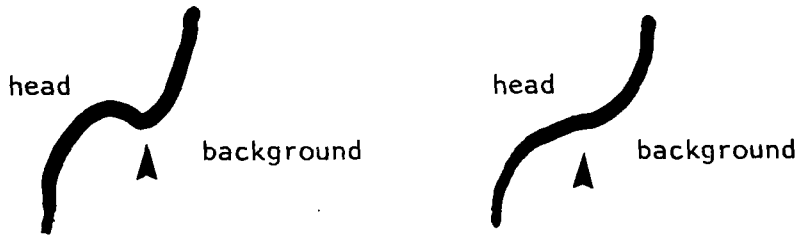


Figure 4. Heads. Note lack of consistancy in size and contrast.



Chin is at turning point or point of least slope.

A B

Fig 3

vertical search looking for a cliff near the point just found. The top, front, and back of the head are found by region growing and thresholding, and the two distances 'head-height' and 'head-width' are noted. For examples of how this turned out in practice, see figure 4. Note the lack of inter-frame consistancy caused by having sepatate cameras take the pictures.

Shoulder Specialist

The shoulder specialist logically follows the head specialist. Given the size of the head, and its position, one can readily make an approximate guess of shoulder position. This is used to obtain a suitable close up view from which to search more closely for the shoulder. The algorithm used is 1/2 of the head-height below the chin in Y, and directly below the back of the head in X. By examining this close up, the shoulder position is found.

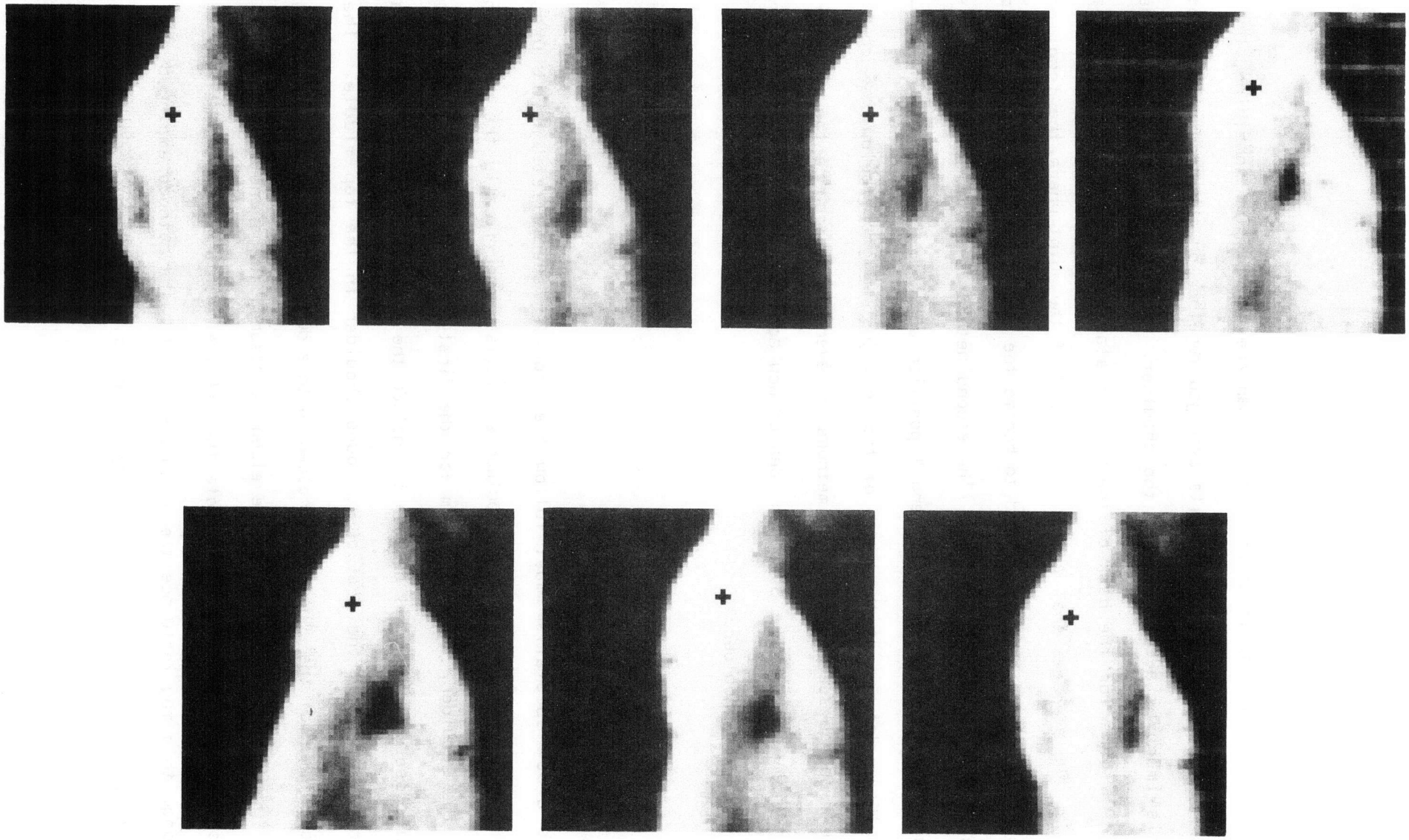


Figure 5. Shoulders.

In general if there is more than one way to find something, both are implemented, and the results averaged together on the assumption that this improves accuracy. For the shoulder, there are two methods. The first looks for the back of the neck to start curving outward from the body, and averages that Y value with the Y of the point where the front of the neck starts curving forward to become the chest. The X value if found by centering within the body. The second method is simply a distance of 'neck-width' below the chin, and X position centered in body. The neck-width is defined as the width of the neck just slightly below the chin. The final guesses of these two methods are averaged to produce the shoulder position. See figure 5 for an idea of how accurate these methods are in finding the shoulder.

Elbow Specialist

The elbow specialist must follow the shoulder specialist, as it uses information gathered by the shoulder specialist, specifically the position of the shoulder. The algorithm for the first step (finding a window for the close-up which will be sure to include the elbow) is to view through a window that has the previously found shoulder in the top middle. The recognition of the elbow also proceeds with a dual method attack, averaging the final positions found. The elbow is assumed to be one 'head-height' distance from the shoulder. Note that if the shoulder or head-height is wrong, this may influence the elbow specialist. The methods employed are: 1) to pick amongst the points along the arc defined by the shoulder point

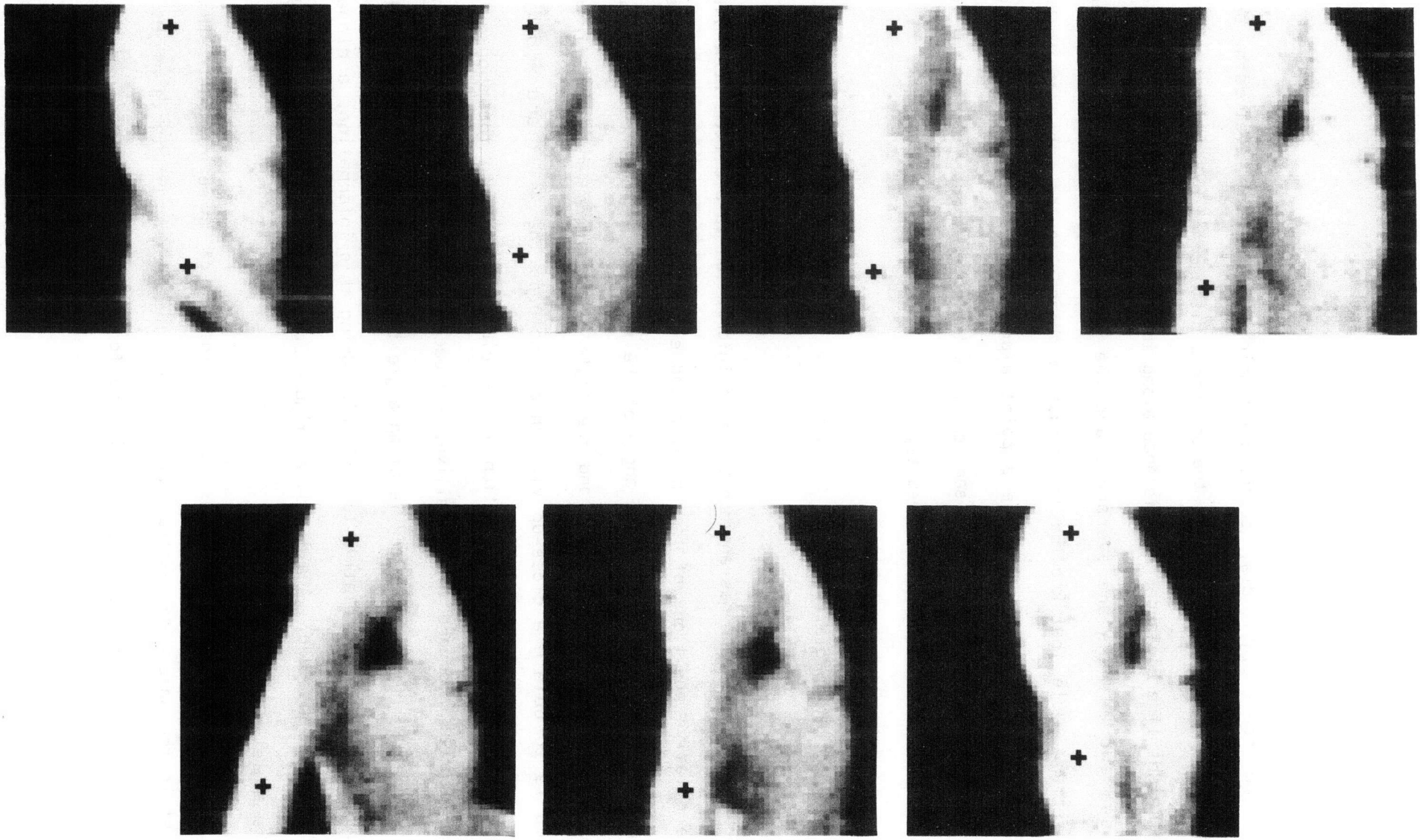


Figure 6. Elbows.

and the 'head-height' distance the point which has the most level path, and 2) to follow the ridge that is the arm for a distance of one 'head-height' (straight line distance, not distance along the path). The point chosen in these two methods are averaged, and the result is termed the elbow position. Note that knowing the structure, hence knowing where to look, (constraining the elbow to be a point along a specific arc), greatly simplifies the task. Figure 2 earlier shows the two methods graphically in A and D. Figure 6 shows some results,

Wrist Specialist

The wrist finder now has another piece of information to aid it. It knows the relative positions of the shoulder and elbow, hence the angle of the upper arm, which constrains the angle of the forearm. Using the knowledge of the position of the elbow, and the angle of the upper arm, the first step is to obtain a close-up view which includes the wrist and elbow. Calling the straight down position of the elbow relative to the shoulder zero degrees, forward swing positive, and backward swing negative, it can be seen that if the upper arm is at an angle theta, the forearm must be at least theta degrees relative to vertical. Using this information, a close up view is obtained. Again look at figure 2b. The arc is costrained by this angle information.

The wrist is found by using a combination of three methods. The first is the same ridge follower used to find the elbow. It starts at the elbow, and follows the ridge that is the forearm for a distance of one

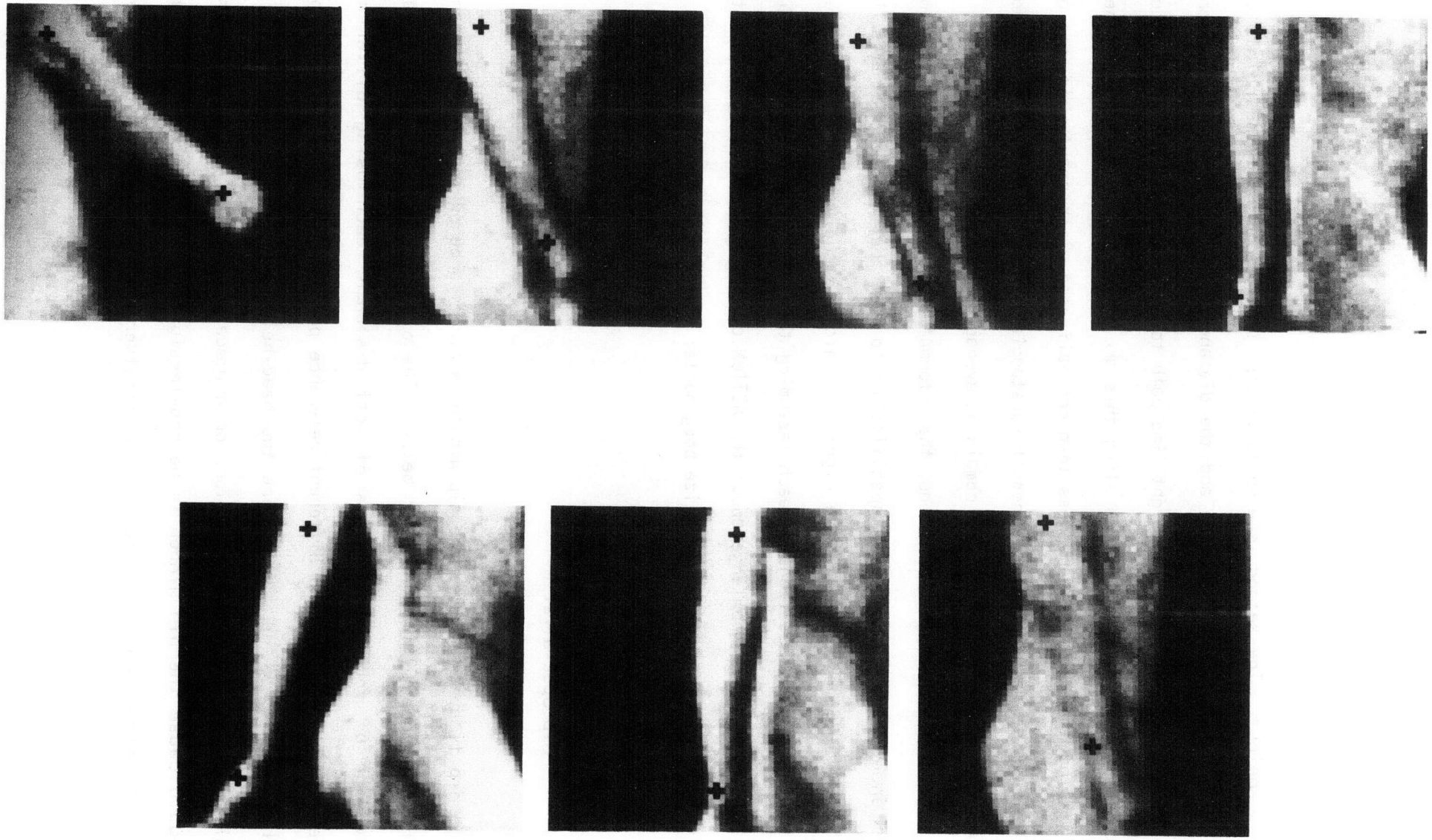


Figure 7. Wrists.

head-height. The second is also familiar; it picks amongst the points on the arc defined by the elbow and the distance one head-height, the point that has the most level straight line path to the elbow. The third method also picks amongst the points along this arc, only it looks for two lines crossing this arc that are less than one half head-width apart. If any of these three methods gives a result substancially different from the other two, it is discarded, and the remaining averaged.

After the wrist is found, the information found in the arm position is encoded for later use, specifically to aid the leg specialists in determining which leg is the right and which is the left, and provides clues as to the positions of each (assuming the subject is walking). This is an example of knowledge about the ACTION of the subject as well as its STRUCTURE being used to recognize body parts. See Figure 7.

Hip Specialist

The hip finder is next. The hip window is directly below the shoulder, and the only question is how far below. The distance used is actually an average of four computations of that distance, each using different measuring distances and different reference points to measure from on the theory that if the window were two head-heights below the shoulder, a mistake in the shoulder position, or a miscalculation of the head-height would cause the hip window to be hopelessly off. The computation used involves head-height, head-width, shoulder-to-top of head distance,

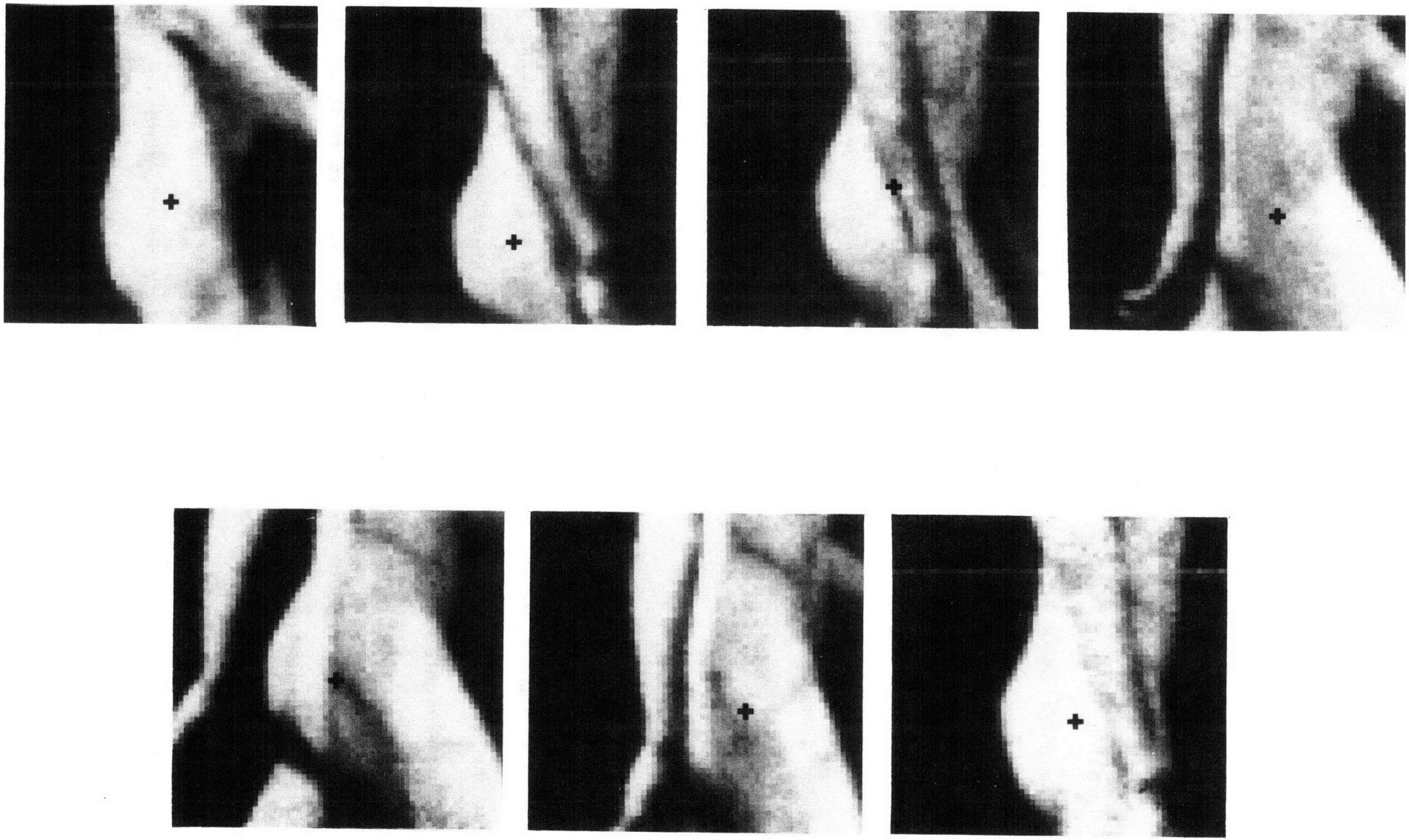


Figure 8. Hips. See also figure 2c.

shoulder reference point, and center-of-head reference point. The net effect is to obtain a window which includes the hip.

The roundedness of the subject's rear end, if unobscured, is used in locating the hip. Lines are drawn perpendicular to the back profile, and the point where most of the lines intersect is used for the Y of the hip. The X is centered within the body. This can be seen in figure 2c, where the perpendictular lines are shown superimposed on the picture. If the subject's back profile is obscured, the wrist position (since it must be hanging near the hip in order to be obscuring) and the previously computed distance below the shoulder are used to find a suitable guess of Y position for the hip. The X is again found by centering within the body. See figure 8.

Knee Specialist

The knee is found by following a ridge down from the hip for a distance of about one and one half head-heights (actually (3/4 hip-shoulder distance + 1 1/2 head-heights) divided by 2), and by examining an arc of that length from the hip. The information gathered on the position of the wrist is used to set the options of the ridge follower, i.e. to hint at the direction at which to go. For example, if the right wrist is behind the body, the right leg is searched for in front of the body. Also if the arc finds two legs, the position of the right arm (closest to cameras) is used to determine which leg is the right leg. The point found is then "verified", by looking for a bend in it, then centered within the leg. Not

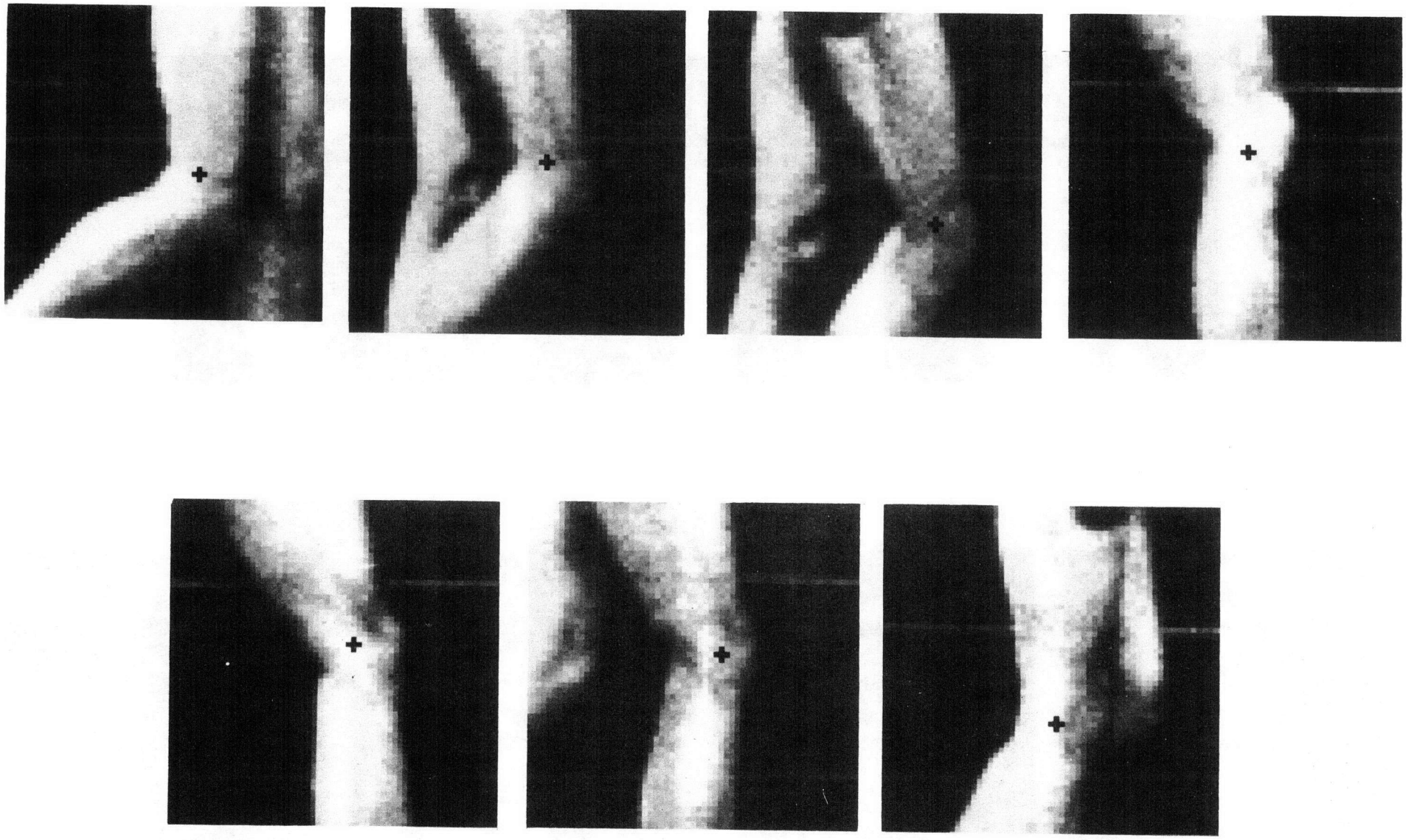


Figure 9. Right Knees.

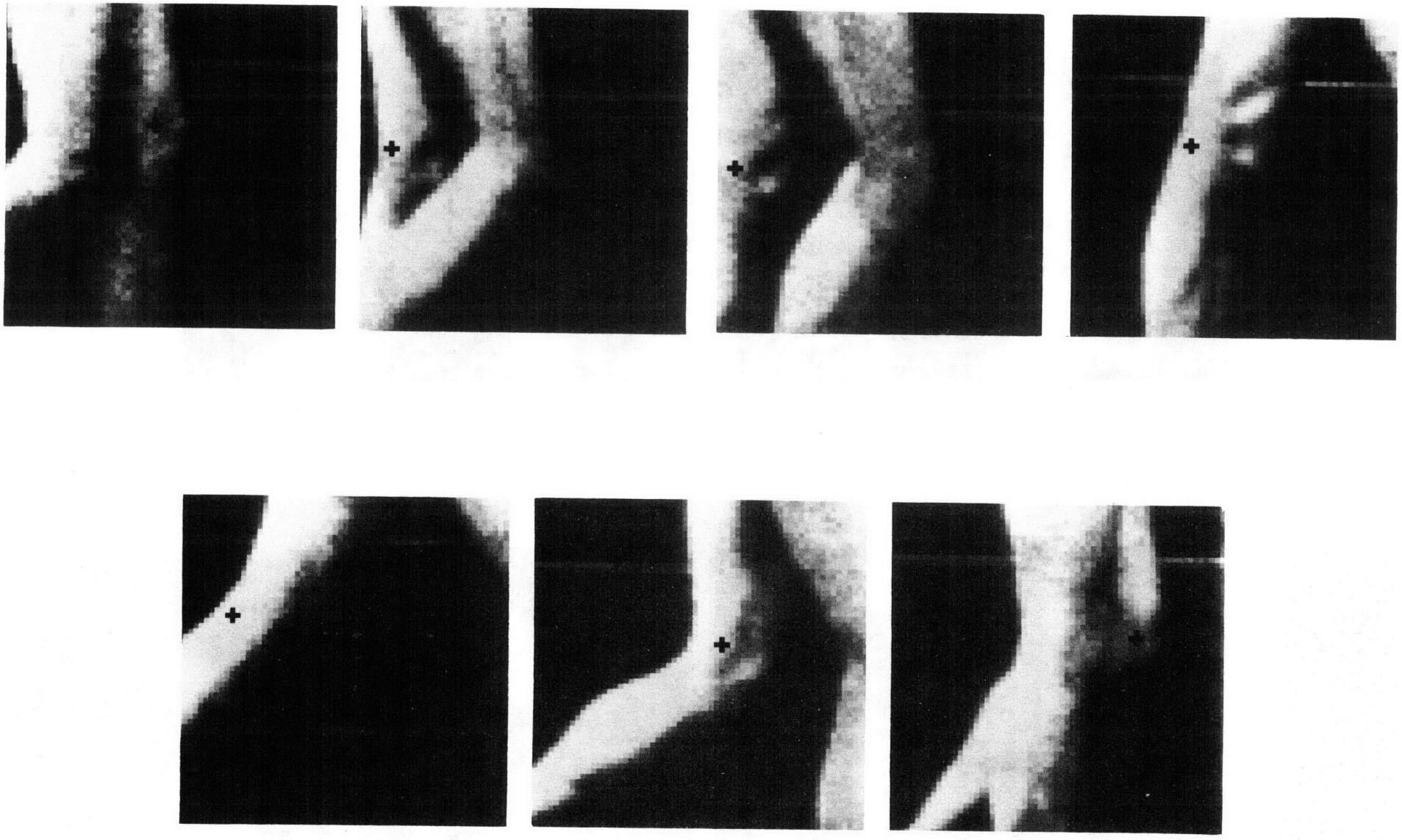


Figure 10. Left Knees.

all knees are bent, but then even a human has difficulty pinpointing the position of a straight knee viewed from the side, and can only use its distance from the hip and ankle. To find the second knee, the hip-first knee distance is used to narrow the search to only those points near the arc this distance from the hip. These are searched, and the knowledge of the first knee's position and the knowledge of whether the second knee should be in front or behind it is used to pick out the second (left) knee. This position is also "verified" by looking for a bend near it, and modified if necessary, then finally is centered within the leg. Figures 9 and 10 show the knees found.

Ankle Specialist

The foot specialist is actually an ankle specialist. It follows the ridge of the leg down from the knee (can be used from either knee), for a distance of 'hip-knee distance' minus one half head-width. This is a little short of the ankle. It then swings an arc centered at the knee through this point to find the center of the leg and its width. The radius of this arc is gradually increased until the ankle is detected as a sudden increase in the leg width. This ankle specialist can be applied to either leg, but when the right leg intervenes between the left knee and left ankle, it has problems. This can be dealt with through the use of other techniques, but this was not done here. This specialist should be two separate specialists, one for each ankle. See figures 11 and 12. Note that

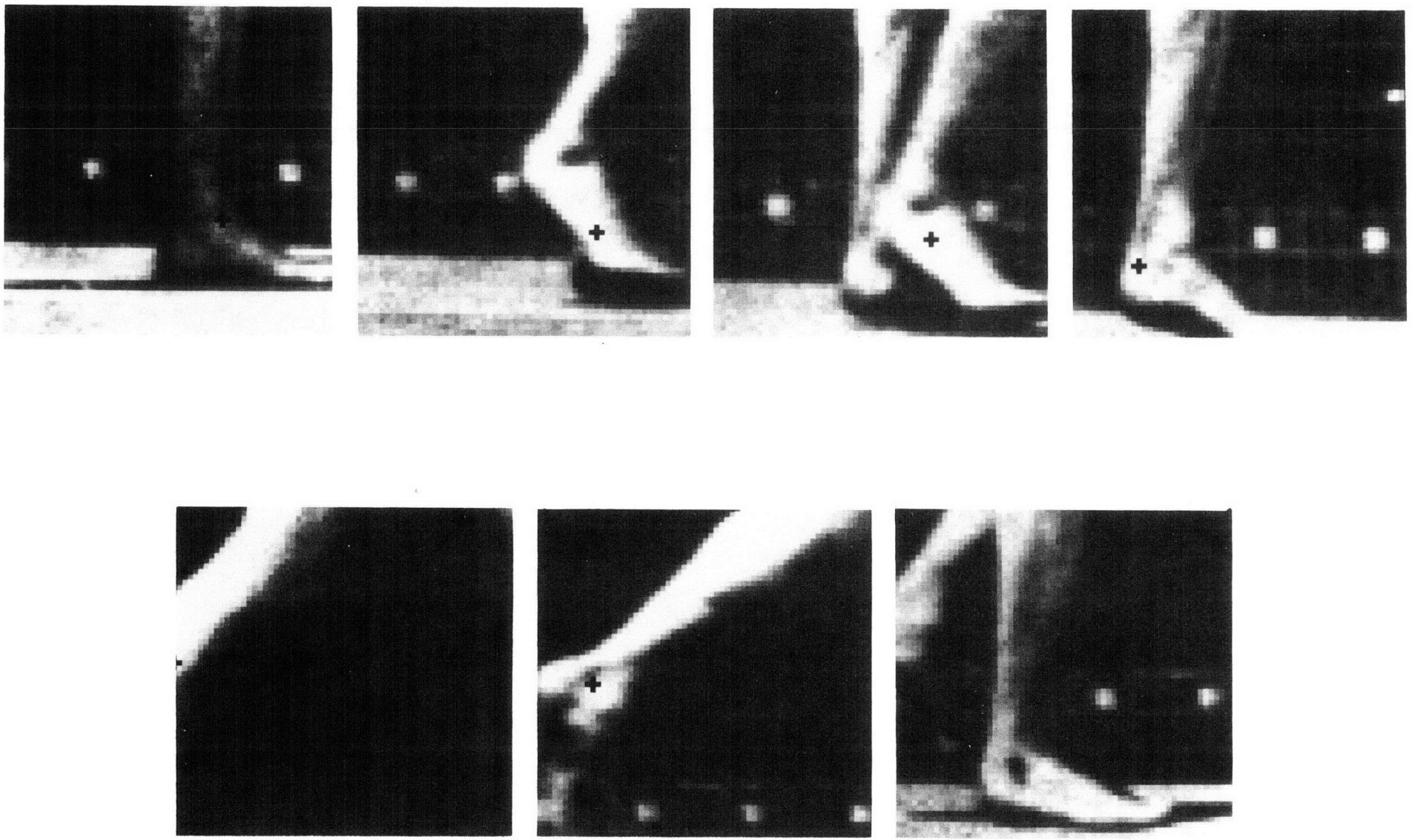


Figure 12. Left Feet. Program knows that the fifth picture does not contain foot.

in figure 12 the fifth foot is not visible. By referring to figure 1, it can be seen that the back foot isn't there at all. The program recognized this fact and returns its best guess.

Results and Conclusions

As can be seen thus far, for the series used, the specialists seem to do their jobs with reasonable precision. The obvious question is how good are they in general. To this end, I viewed a picture from another series in Muybridge's collection of persons walking. The other series are not as good quality pictures for various reasons, and the fact that the methods do not work as well is not surprising. The fact that they are able to recognize body parts from sub-optimal pictures at all shows the generality of the methods. See figure 13. As can be seen, there is moderate success, but again the foot specialist can be improved.

In conclusion, it can be seen that the idea of the specialists seems to work reasonably well, and would work better if there was more redundancy employed such that the failure of any one specialist would not cause severe hardship to the other specialists. The more knowledge available to the specialists, the better they can function. The knowledge used here included linkage lengths, structure (what is attached where), constraints on joint angles, and hints based on knowing how a person is likely to move when he is walking. The top directed use of tools seem to vastly improve their usefulness. Given more frames, and greater consistancy between them, knowledge of the previous positions could be used to compute velocities and accelerations of points about other points, and would be a good predictive tool. Thus as a method for actually analysing

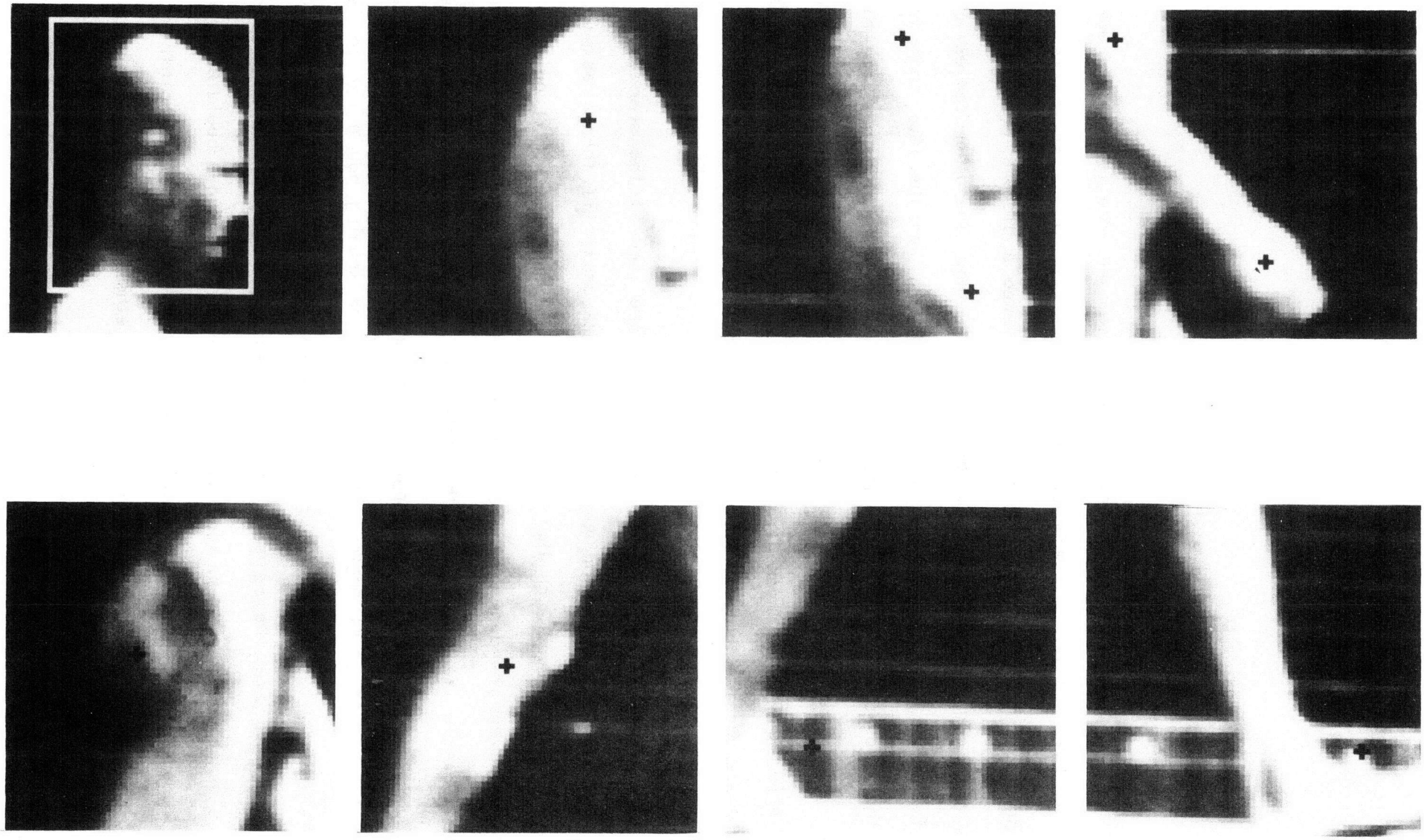


Figure 13. Alternate subject viewed to test generality of methods.

motion, direct visual analysis via computer seems to lack behind other methods, but as can be seen from the pictures, top down methods combined with a strong internal model can recognize human body parts from a photograph.