Automatic Visual Recognition of Armed Robbery

Jaime Dever, Niels da Vitoria Lobo, and Mubarak Shah
Computer Science
University of Central Florida
Orlando, Florida, USA
E-mail: {jdever,niels,shah}@cs.ucf.edu

Abstract

We propose a method by which to analyze silhouettes and recognize a classic holdup position of armed robbery. In such a situation, one actor levels his or her arm while another actor raises his or her arm(s) into the air. The core of this algorithm is skeleton analysis. We attempt recognition by first segmenting the skeleton of the silhouette into separate pieces of the body, then identifying the positions of the arms. We show that our algorithm correctly utilizes skeletons to identify parts of the human body and recognize these holdup positions.

1. Introduction and Related Work

Video surveillance has long been used in an attempt to prevent crimes by providing a ready means of identifying the perpetrator and ensuring that he or she is held accountable. Useful as this may be, it remains a passive method of crime control. Computer recognition of activities in such situations could thrust video surveillance into an active role by allowing police to be alerted automatically, hopefully in time to prevent loss of life or property.

Recognition of human activities is already the subject of much research. In one such project involving surveillance of parking lots[5], characterizations were based solely on the motion of the objects in the image. The movements were then compared to a set of possible events.

In another project[1], the trajectories of a set of objects were compared against models for specific activities. The actors themselves were not investigated; they were only tracked and their vectors analyzed. For example, football players were viewed from above and their motions were compared against a database of plays.

Collins[3] proposed a means of identifying individuals by the appearance of their silhouettes while walking, in a full frontal or back camera view. Silhouettes from key frames were compared to known walking patterns previously recorded. Wren and Pentland[7] proposed a robust, 3-D skeletal model for tracking people and recognizing purposeful motion. Two cameras were used to get the 3-D information, and motion not explained by physics was said to be purposeful. Guo[4] used a skeleton model based on the 2-D information provided by a single camera. Rather than calculating and inspecting the silhouette's skeleton, this algorithm seeks to fit the skeleton model to the silhouette. The entire person was assumed to be visible, and the silhouette to be very good. *Walking, running,* and *other* were the activity classes identified.

None of these holistic approaches will identify actions as specific as levelling an arm at another actor that raises its arm(s) into the air in response. In our approach, we investigate the individual motions and silhouette appearance more closely. We propose to analyze the skeleton, or medial lines, of the silhouette in order to identify the position of the arms and recognize the classic hold-up positions of armed robbery.

2. Overview of Algorithm

The steps in our algorithm are:

1. Silhouette Extraction: The people in the scenario must first be identified. This is accomplished through background subtraction [6] together with connected components. To achieve a smooth silhouette for skeleton calculation, this is followed by dilation and erosion.

2. Skeleton Segmentation: First, the skeleton is calculated. Points of interest (POIs) are then determined and used to find the individual segments of the skeleton.

3. Segment Identification: Position, length, and slope are then used to identify the torso, legs, arms, and head, if they are present.

4. Arm Analysis: Finally, the slopes of the arms are used to determine whether a possible armed robbery exists in the frame or sequence.

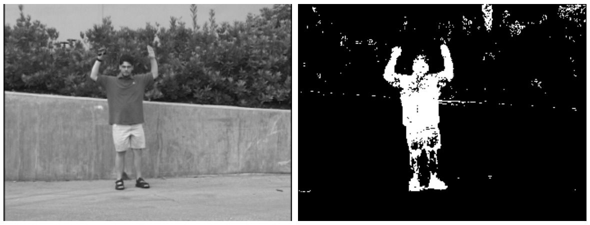

Figure 1. Original image and labelled foreground.

3. Silhouette Extraction

Simple identification in frames and tracking through a sequence is accomplished using Stauffer's background subtraction[6] to label foreground pixels (Figure 1) and connected components. In this implementation, the initial background must be known and the camera must be fixed. Canny[2] edges help to simplify size thresholding of blobs, limit the interference of shadows, and ensure that previously acquired silhouettes are not culled unnecessarily due to size restrictions.

The first connected components phase associates all canny edge pixels (Figure 2) of the foreground with nearby edge pixels also in the foreground. Valid objects tend to have more definite edges than do areas of noise; therefore the size difference in number of edge pixels is often greater than in total number of pixels.

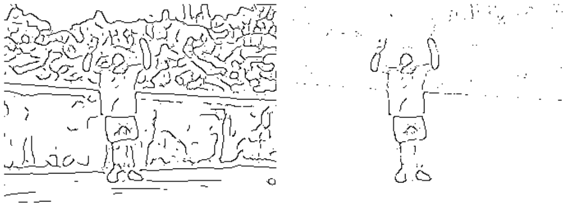

Figure 2. Edges and foreground edges.

The remaining foreground pixels are then grouped by connected components within a bounding box determined by the extremes of the grouped edges. This is to accommodate for shadows. Except in extreme cases, shadows do not tend to yield many edge pixels; all or part of a shadow will then be outside of the boundaries and excluded from the silhouette. In subsequent frames, the previous location of the blob is the basis of the bounding box rather than the edges, after allowing for some movement. Blobs found in previous frames, if still present, should be within the size threshold for new silhouettes.

Traditional dilation and erosion are used to smooth the silhouette for skeleton calculation. This is followed by another erosion in which blob pixels with too few blob neighbors are discarded. Only a basic silhouette like those in Figure 3 is needed for the next steps, as the skeleton analysis is intended to compensate for normal errors in the extraction.

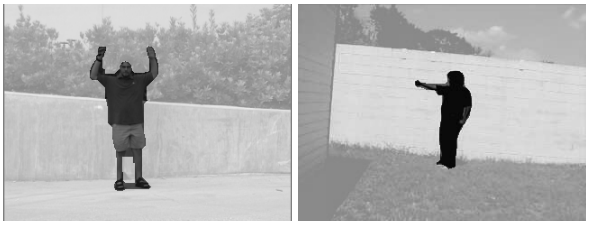

Figure 3. Silhouette.

4. Skeleton Segmentation

The silhouette's skeleton is calculated by eroding the edges of the region until only curves with width of one pixel remain. Then, points of interest (POIs) are identified (Figure 4). The two classes of POI, endpoint and intersection, are determined by the number of neighboring pixels in the skeleton: one neighbor indicates an endpoint of the skeleton, while three or more indicate an intersection between two or more curves in the skeleton. Each POI calculated is actually a cluster of pixels, though they are stored and treated as a single point. Non-POI pixels of the skeleton are then grouped by connected components to make up the individual segments of the skeleton. Their endpoints, lengths, and slopes are calculated for analysis, shown in Figure 5.

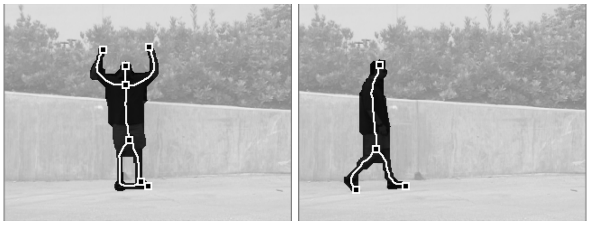

Figure 4. Skeleton; boxes indicate POIs.

5. Segment Identification

Different parts of the body are identified based on the POIs, length, and orientation of each segment. Possible problems with the silhouette are also addressed.

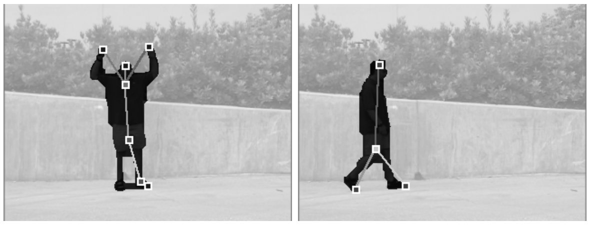

Figure 5. Skeleton curves represented by individual lines.

5.1. Segment Characteristics

It is currently assumed that the bodies are relatively upright. Thus, the torso is nearly vertical and one of the longest. This segment is always present; if no other segments are present, it spans the entire height of the skeleton. The nature of its POIs give clues to the presence of other segments. If the top POI of the torso is an intersection rather than an endpoint, then at least one arm and the head are present. The head is a relatively small extension from the top of the torso. The bottom point of the head corresponds to the top point of the torso as well as the shoulder end of any arm; its top point is an endpoint. Arms also have one intersection with the top of the torso and one endpoint. Legs exist when the bottom point of the torso is an intersection rather than an endpoint, and are the bottommost segments. They almost always occur in pairs, and their top point corresponds to the bottom of the torso. See Figure 6.

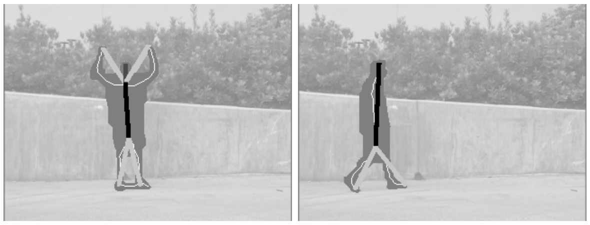

Figure 6. Segment identification (a) 2 arms, 1 head, and 1 torso identified. (b) 1 torso and 2 legs identified.

5.2. Silhouette Compensation

Our algorithm is designed to compensate for a number of shortcomings in the silhouette. Such faults lead to distinct errors in the skeleton.

1. Loops: Occasionally, there appear two segments that share the same endpoints. This occurs either when the legs are joined by shadow at the bottom of the silhouette, or when a hole is found in the silhouette. In the former case, the legs are separated into two segments as shown in Figure7. In the latter, the two segments are joined to be only one.

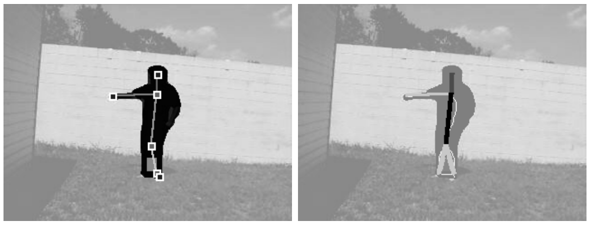

Figure 7. Legs taken from a loop.

2. Spurs: Small, irrelevant segments often occur as a result of irregularities in the silhouette. In instances where one of its ends is an endpoint, the spur can simply be deleted because it is an offshoot of another segment and does not belong. The true segments wrongly separated by this error often need to be joined. In instances where the spur has another segment on both ends, the spur must be added to one of them and it must be determined whether these two segments are also one and the same. This occurs most often when the arms do not meet at the same part of the torso, so the points of these segments must be adjusted as well. One frequent example is ”feet” on the end of leg segments, as shown in Figure 8.

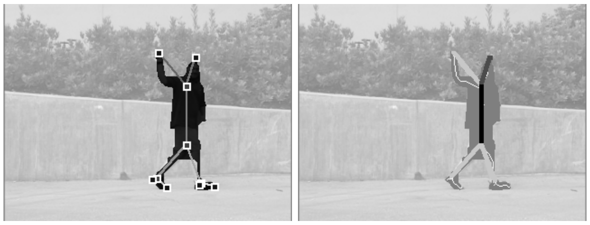

Figure 8. Irrelevant segments removed from legs.

3. Links: An additional segment occasionally joins the bottoms of the legs. The cause is shadow, as with loops (Figure 9). In this case, the erroneous segment can simply be deleted.

6. Arm Analysis

Once the arms have been identified, their positions are analyzed based on their slopes. (We assume the figure to be in the first quadrant of the Cartesian coordinate plane.) A

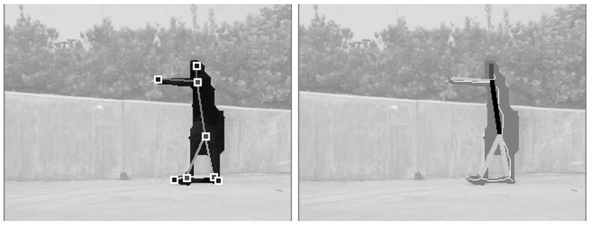

Figure 9. Linking segment removed from legs.

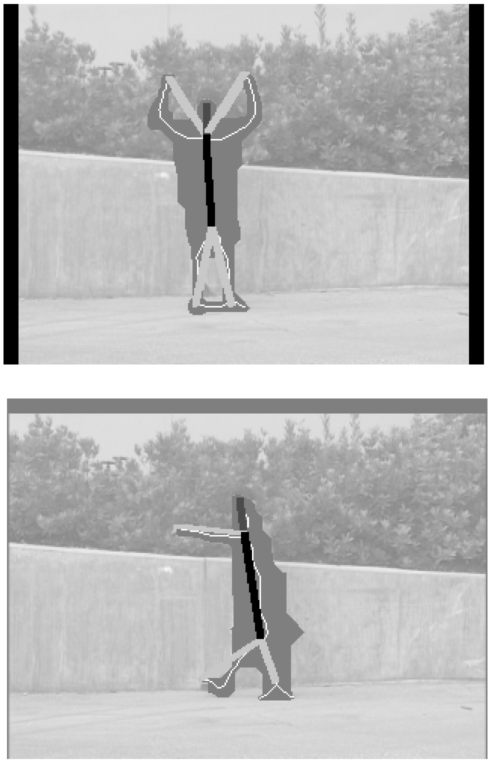

Figure 10. Alarms (a) raised alarm; left line indicates left arm, right line indicates right arm (b) level alarm; bottom line indicates right arm, a top line would indicate left arm

level arm is defined as being within 20 degrees of horizontal within the frame, while a raised arm must be at least 40 degrees above the horizontal. The tangents of these angles are 0.4 and 0.8 respectively, and may be tested against the slope m of each line, given by $y = mx + b$. A level arm is therefore one with $-0.4 < m < 0.4$. When an arm first becomes distinguishable from the torso, it nearly always falls within this range; this is therefore the lowest alarm level. The second level is when the victim's arms are determined to be raised. In this case, a left arm must have a slope $m < -0.8$ and a right arm must have a slope $m > 0.8$. The highest level of alarm occurs when there are at least two silhouettes present and the first two alarms are both recognized.

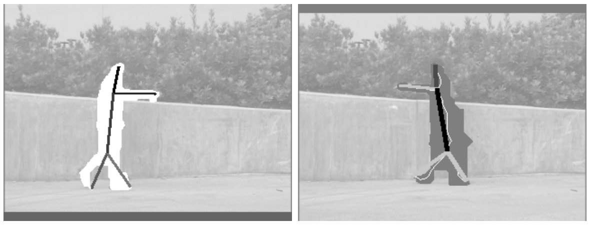

Figure 11. Results from level-arm sequences.

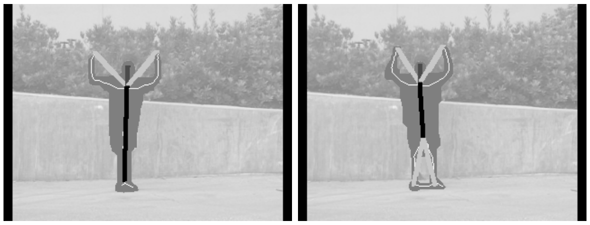

Figure 12. Results from raised-arm sequences.

7 Results

Figure 11 represents output from different sequences in which the actor triggers the level-arm alarm. In these examples, the 5 body segments are correctly identified and the proper alarm is successfully reached. Figure 12 results from different sequences in which the actor triggers the raised-arm alarm. Here, the 5 body segments are correctly identified and the raised-arms alarm is successfully reached. In the second image, the legs are successfully separated and a spur successfully removed. Table 1 shows the success rates over 10 separate sequences. Our algorithm successfully identified the different segments of the body in approximately 80% of the frames overall . In all of the frames where the arms were properly segmented the algorithm produced the expected alarm.

8. Discussion and Future Work

Our algorithm analyzes the skeleton of a silhouette to determine the positions of the arms. This information indicates whether any or all of the possible criteria for a classic holdup situation are met in the frame. The algorithm identifies segments of the body with a high rate of success, compensating well for poor silhouettes. However, results

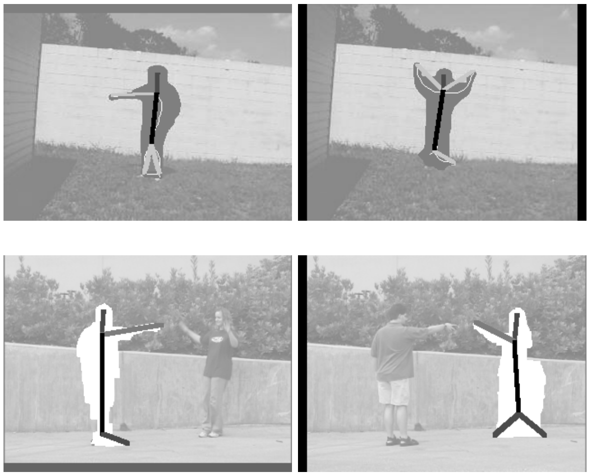

Figure 13. Results from additional sequences.

sequence name	frames	successful id
WALTAIM1	60	78.3 %
WALTAIM2	42	85.7 %
WALTAIM3	63	69.8 %
WALTAIM4	71	81.7 %
WALTUP1	60	53.3 %
WALTUP2	71	63.5 %
WALTJME *	189	56.6 %
MIKEAIM1	262	93.0 %
MIKEAIM2	180	92.2 %
MIKEMULT	241	89.2 %

* There are two persons in this sequence; each is considered individually for this calculation.

Table 1. Results from 10 sequences

are still sensitive to errors in the silhouettes; the sequences with lower success rates did have silhouettes of lower quality than the sequences with better rates. For instance, the MIKE sequences generally yielded better results than did the WALT sequences. In the latter set, the background was much less stable and provided less contrast than in the other, and the actor moved very quickly.

The algorithm is view invariant within obvious limitations. It is not important whether the front or back of the actors are facing the camera; nor do they need to face the camera directly. As shown in Figure 14, rotation is acceptable to a large degree. However, the arms must be visibly separate in the silhouette from the torso. Thresholding and data from past frames could improve this performance slightly, but at the expense of some false alarms in other cases.

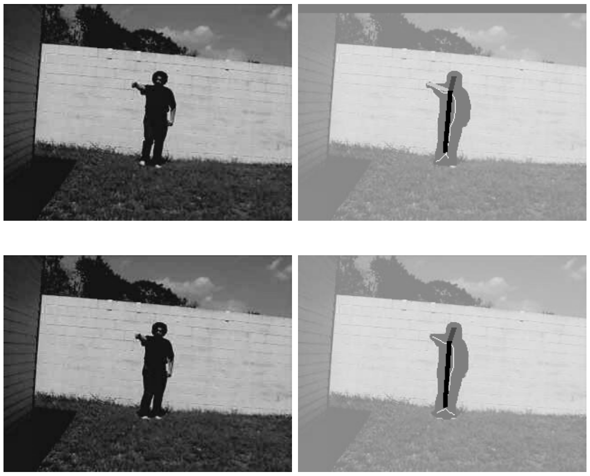

Figure 14. Lose arm in rotation.

Highly irregular blobs result in spurs, connections, and other extraneous information that can confuse the analysis if too extensive. It is suggested that silhouettes yielding inordinate numbers of POIs and segments should undergo additional smoothing, then be addressed again by the skeleton analysis algorithm. Additional work can be done with the compensation portion itself. Alternative methods of identifying the head are being pursued in order to increase the accuracy of overall identification. Additional alarm criteria to narrow the recognition of this situation as well as accept other situations are also worthy of attention.

References

[1] A. Bobick and S. Intille. Recognizing planned, multiperson action. *Computer Vision and Image Understanding*, 81(3):414–445, 2001.

[2] J. Canny. A computational approach to edge detection. In *IEEE Transactions on PAMI, 8(6)*, pages 678–679, 1986.

[3] R. Collins, R. Gross, and J. Shi. Silhouette-based human identification from body shape and gait. In *5th International Conference on Automatic Face and Gesture Recognition*, May 2002.

[4] Y. Guo, G. Xu, and J. Tsuji. Understanding human motion patterns. In *12th International Conference on Pattern Recognition*, volume 2, 1994, pages 325-330.

[5] Y. Ivanov, C. C. Stauffer, A. Bobick, and E. Grimson. Video surveillance of interactions. In *Proc. of the CVPR'99 Workshop on Visual Surveillance*, November 1998.

[6] C. Stauffer and E. Grimson. Learning patterns of activity using real time tracking. In *PAMI, Volume 22*, pages 747–757, 2000.

[7] C. Wren and A. Pentland. Understanding purposeful human motion, 1999.