Multi-View Human Pose Estimation using Modified Five-point Skeleton Model

Daniel Chen, Pi-chi Chou, Clinton Fookes and Sridha Sridharan
Image & Video Research Laboratory
Queensland University of Technology
GPO Box 2434, 2 George St
Brisbane, Queensland 4001
{*daniel.chen*, *p.chou*, *c.fookes*, *s.sridharan*}@qut.edu.au

***Abstract*—This paper examines the task of estimating the 3D pose of a human subject acquired from multiple views within a multiple camera surveillance network. We utilised a modified five-point skeleton model with potential application in human action recognition and gait recognition. This paper proposes automatic initialisation and recovery of human pose. Feature tracking and motion prediction are incorporated to increase the accuracy and the robustness of the model. Although the model is tested within the area of video surveillance, it has the potential to extend to other areas such as Virtual Reality, content based retrieval and compression of video. The proposed algorithm is evaluated with the IXMAS database and is demonstrated to produce promising results for 3D pose estimation from a multi-view camera network. Outcomes are also evaluated using feature trackers.**

I. Introduction

Automatic human pose estimation is an extremely complex challenge facing the computer vision community. The ability for a computer to automatically decompose and estimate the structure and pose of the human body has various applications such as human action recognition [1], gait recognition [2], and virtual reality [3].

Human action recognition is performed by classifying parts of a video sequence as a particular action. Systems need to be 'taught' how to recognise actions by first showing many examples of that particular action. Algorithms used for recognition include template matching, hidden Markov models (HMM), Bayesian networks and neural networks, along with variants of these such as coupled HMMs and time-delay neural networks [1]. HMMs and its variants are by far the most common techniques currently used.

Gait recognition is an identification process that extends action recognition and is often performed on walking or running motions. Gait recognition can be categorised by two distinct approaches: appearance-based and model-based approaches [4]. The appearance based approach performs gait recognition directly from the human silhouette. In general, feature extraction for this approach is a simple process. The common features include the prediction of position, velocity, shape, texture and colour. These approaches are often limited when dealing with large view variations, suffer from view dependency, and are sensitive to changes in lighting, occlusion, noise, and clothing. Model based approaches involve complex model fitting and tracking frameworks. Constraints such as symmetry and degree of freedom (DOF) are often employed to minimise the computational cost. Common features include the angles between body parts, and static length of the body parts. These approaches are often not view dependent, but suffer from poor tracking of the upper limbs due to self occlusion and inconsistent movement.

A difficult issue still requiring significant development is the robust estimation of pose in 3D which can be acquired from multiple views. Current methods are often limited to a single view, are constrained by view-dependent algorithms which cannot operate on any arbitrary view, or they lack the ability to effectively combine the required information from multiple views. Recent work in this area is targeting these issues and is summarised in the following section.

With the goal of human action recognition and gait recognition in the area of surveillance we are seeking a technique that is simple to implement and has fast execution time for real-time application. The reason for a multi-view approach is two fold; to allow accurate 3D reconstruction and to minimise the effects of occlusion. The modified five-point tracker has this potential for pose recovery. We have implemented the tracker using multi-view video sequences and construct a 3D model to estimate human pose. Results illustrate the model is able to recover pose using simple decision making rules.

This paper is outlined as follows. Section II provides an overview of some related work in the field. Section III presents the automatic pose estimation technique. Section IV presents some experimental results which have been tested on the IXMAS dataset [5]. Finally the paper is concluded with a discussion and an overview of future work in Section V.

II. Related Work

One of the simplest approaches for recovering human pose within the 2D domain is often referred to as 'star' skeletonisation. This approach traces the contour of the silhouette of a person, calculating the distance to the centroid of the silhouette and taking the five most prominent peaks of this distance measure. These five points correspond to the head, hands and feet, and when connected to the centroid, resembles a star shape. Fujiyoshi and Lipton [6] use the lower two points, or

the feet, to distinguish between walking and running from its periodic frequency. Slight adaptations to the algorithm have since been made and utilised for other applications, such as fence climbing detection [7] and also human action recognition [8].

A more complex method which also labels the elbow and knee joints, the hip and neck has been proposed by Thome et al. [9]. A silhouette of a person is first skeletonised using a method based on Voronoi diagrams and then the skeleton is pologonalised. The skeleton is then decomposed into a directed acyclic graph and matched against a human model to label the points in the skeleton. Tracking was applied to each individual point to help resolve ambiguities when matching and to assist in dealing with self occlusion.

The Pfinder system [3] tackles the pose estimation problem by modelling the human body as a series of 2D blobs, with each corresponding to parts of the human body (head, hands, feets, upper and lower body). It uses silhouettes extracted from motion detection. The system has been used in various implementations, such as a control mechanism to navigate within a 3D virtual game environment, and was also the basis for the American Sign Language recognition system.

Ren et al. [10] computes the edge map of the image and then identifies parallel lines from the map. Pairwise constraints between body parts are then used to assemble a human body from these lines. Mori et al. [11] also uses an edge map for their implementation. A normalised cut approach is then used to segment the image with each of the segments passed through various detectors in an attempt to identify limbs and the torso.

The methods presented so far all deal with 2D reconstruction of pose. 3D approaches have also been explored. Peursum et al. [8] use multiple cameras to expand the star skeletonisation algorithm into 3D. The standard 2D approach is performed on each of the available views, and then the detected points are matched, given the relative positions of the cameras, to generate 3D locations for the points.

Agarwal and Triggs [12] used an example-based approach to recover 3D pose from monocular video. Nonlinear regression was used to learn mappings between silhouette shape descriptors to a 3D pose. Sminchisescu and Triggs [13] used image matching, joint limits and non-self-intersection constraints to implement their 3D recovery system. Gavrila and Davis [14] use a simple volumetric 3D model to represent the human body. An iterative process matched the outline of the 3D model to an edge map generated from the input image to estimate pose.

As illustrated from the literature, the model based approaches generally involve complex model fitting and tracking frameworks. These methods are often computationally expensive and difficult to implement. The advantage of model based approaches however, is their view independency and robustness when dealing with occlusion. On the other hand the non-model based approaches are relatively simple to implement with faster performance, but they heavily rely on the viewing angle and often fail when dealing with occlusions. We have chosen the modified five-point model due to its simplicity and the potential ability to overcome issues such as viewing angle dependency and occlusions.

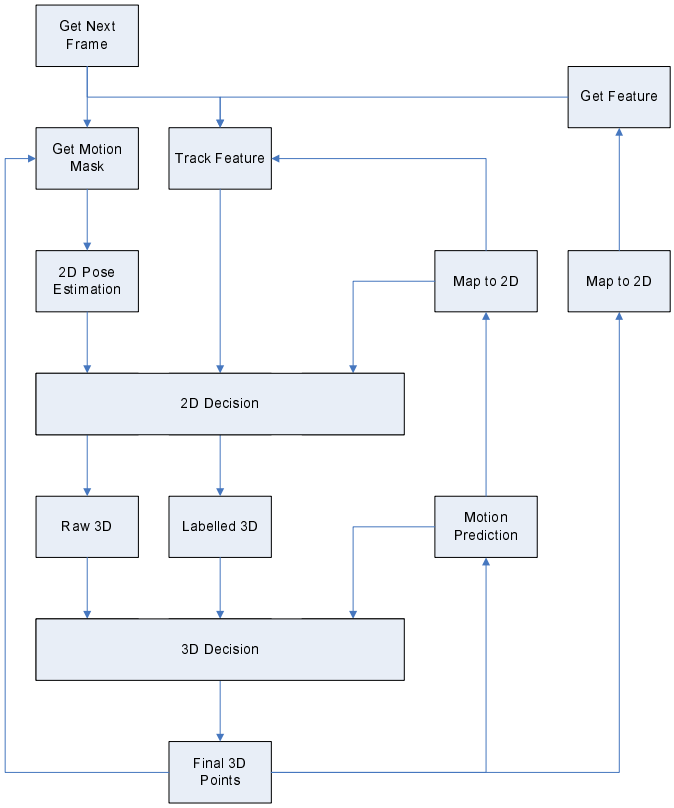

Fig. 1. Flow chart of algorithm

III. Pose Reconstruction

A. Algorithm Overview

We attempt to reconstruct the human pose in 3D by performing 2D pose reconstruction independently in multiple views, and then projecting them into 3D space. Motion detection is employed on the raw image data to obtain silhouettes of the subject. 2D pose reconstruction is then applied on these motion masks. The result of the 2D pose reconstruction is then used to reconstruct in 3D. Velocity based motion estimation is used on the 3D points to help with tracking between frames. Feedback is provided back to the 2D stage by projecting the final 3D results and the motion estimation back into the different views. Feature tracking is performed on the projected final points. This process is illustrated in Figure 1. Each of these stages is outlined in the following subsections.

B. Star Skeletonisation

The 'star' skeletonisation algorithm proposed by Fujiyoshi and Lipton [6] was used as the basis for the pose estimation in the 2D stage. This method involves tracing the boundary of a silhouette and calculating the distance to the centroid of the shape. Extremities are found by finding the maxima in these distance values. Taking the five most extreme points results in

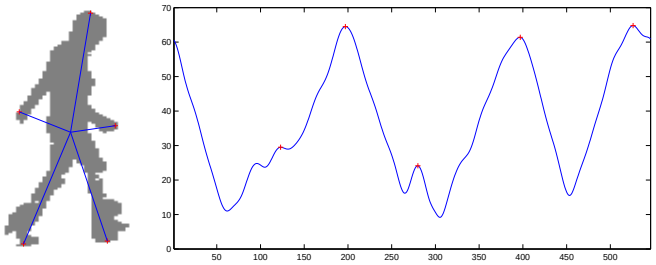

Fig. 2. Basic ‘star’ skeletonisation example

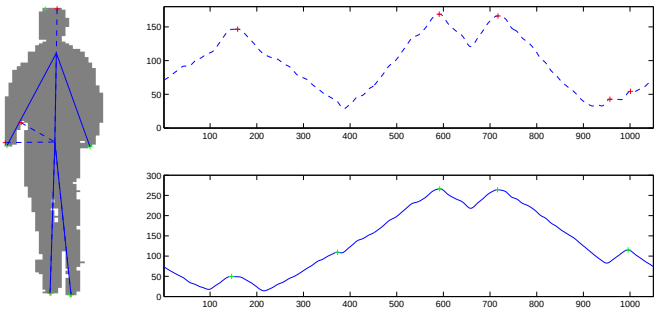

Fig. 3. Comparasion between original and modified algorithm, taking the 5 most prominent peaks

a ‘star-like’ skeletal representation of the human body when joined to the centroid, as can be seen in Figure 2.

Alterations to this algorithm have been made by Peursum et al. [8] and Yu et al. [7]. Yu moves the ‘centre’ point from which the distance is calculated from the centroid to the head. Peursum also moves this centre point, this time to two thirds the way to the head from the centroid. This change corrects a problem faced with the original algorithm in which incorrect points (Figure 3) are detected. In the case with Peursum, the point also approximates the neck/shoulder and serves as an anchor point for the arms.

The approach followed here resembles that by Peursum. The distance to the boundary from the centroid is first calculated and then smoothed slightly through the use of Fourier descriptors. Extremities are found by searching for local maxima. First, a temporary head point is located by taking the topmost point. A point two thirds from the centroid toward this head point is then found, and a second set of extremities are then found based on this point instead of the centroid.

The subject is assumed as upright, and thus the head can be taken as the uppermost extremity point. However, this does not necessarily correspond as the highest point in the silhouette as no assumptions are made on the camera locations. To identify ‘up’ in each of the views, the centroid is first identified in each silhouette and its 3D location is estimated. An arbitrary point directly above the 3D centroid is calculated and projected back into each of the views to obtain a vector pointing ‘up’.

The head is re-identified by taking the topmost extremity from the second set. As each new point is labelled, extremities that are close are removed from both sets. The legs are identified by locating the lowest 2 extremities detected that lie within the lower 1/3 of the silhouette, with the set of extremities calculated from the centroid given preference over the other. The hands are found by taking the two further points remaining from the second set. Should less than two points be found, the remaining hand points are found by taking the closest points in the first set.

No attempt at distinguishing between left or right hands or legs is ever made. Each view is processed independently for this stage so no correspondences are made between the different hands and legs between the views so far.

C. Tracking

Feature tracking was was also incorporated, along with the reprojected motion estimates, to track points as the subject moves between frames. Two different feature tracking techniques, namely local zero-mean normalised cross correlation (ZNCC) [15] and the KLT feature tracker [16], were tested to improve the accuracy and robustness of the system. It was shown the cross correlation technique is more robust in comparison to the KLT tracker in our particular application. This was due to the KLT tracker proving more adept at tracking corner points while the tips of the human limbs are often blurred and the tracking points are easily lost. The cross correlation technique overcame the problem by not tracking single points, however, it is still sensitive to the selection of the centre point for the feature window. Generally the points are initialised to locate the tips of the limbs, and the edge feature will dominate the tracking result therefore it is often lost due to background edges.

A simple motion estimator based on previous velocity information was applied on the final 3D points. These were mapped back into 2D to help with the tracking process. Correspondences between the correct hand/foot are now able to be made due to tracking.

D. 3D Reconstruction

Given calibrated cameras, one is able to map a point in 3D space to a coordinate in the camera view. To do the reverse, at least two different views are need. A point in a 2D view projects into a line in 3D space. Two corresponding points in two different views will produce two intersecting (assuming perfect accuracy) lines, where the intersection is the points’ location in 3D.

For every point of a given ‘type’ (head, hand, etc) we triangulate with every point of the same type in another view, calculating the point where the two projected lines are closest and working out the separation. We then find correspondences between points in different views. This is done by isolating pairs of points with the smallest distance between them at their closest approach. This way, we are able to discount false detections in the 2D stage. It also allows us to use multiple points, eg. having say 4 candidate ‘hand’ points being extracted from 2D to hopefully include correct points that may otherwise be missed. This however, allows for the chance where two unrelated points happen to provide a good correspondence.

This process is applied to every possible combination of two camera views, resulting in numerous detected points in space for each body part. Points are grouped by joining with its nearest neighbouring point and the average of the closest two points become the candidate point for that group. The candidate points with the smallest separations are then labelled as the final 3D points.

This 3D process is also applied again with the distinction between the different hands and feet, giving a second set of 'final' points. This is combined with the first set to produce a more robust result.

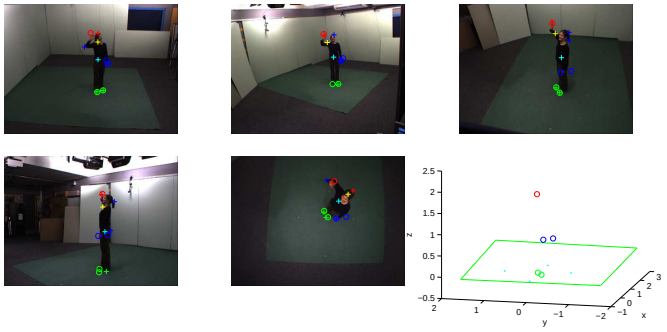

Fig. 7. Failure example

IV. Experimental Results

The objective of our experiment is to test the accuracy and usability of the modified five-point skeleton model for human pose estimation. The algorithm was implemented in Matlab and tested using the IXMAS database [5], which contains multi-view video sequences from five calibrated cameras in a controlled environment. The skeleton was extracted using the silhouettes that where provided with the database. Other motion detectors were also tested and resulted in similar silhouettes. The results presented here are generated without feedback from feature tracking as the feedback process often introduced significant problems during periods of excessive movements and did little to improve results during other periods of normal motion.

An example frame is shown in Figure 4 with the five synchronised camera views visible with a sixth image showing the reconstruction in 3D. The labels are colour coded, with red, blue and green corresponding to the head, hands and feet respectively. The crosses in the camera views show the raw detections from the 2D pose estimation, while the points in the 3D view shows the results of the point projections. The blue crosses represent the centroid and yellow corresponds with the point where the second set of extremities is calculated. The circles represent the final 3D locations. For this particular example, the legs were joined to the centroid with the hands connected to the secondary point to form a more human like stick figure.

Figures 5 and 6 illustrate the results on two sequences; one showing the subject raising her arm to check her watch, while the other shows her walking in a small circle. It can be seen that the system is able to provide reasonable simple reconstruction of a person's pose. Some problems can be seen however. For example, in the watch checking sequence, tracking of the left hand shifts to the elbow as the hand occludes with the body and the elbow becomes more prominent. In the walking sequence, incorrect 3D locations were detected in a few of the frames. It can also be noted that the foot locations detected are significantly lower than where they should be, appearing to be below the ground. This is attributed to the effects of shadows causing problems with the motion masks.

Figure 7 shows a case where the system fails. In this case, our assumption that the head is the upper most point is broken and the result is incorrectly labelled. The system is also unable to handle large amounts of occlusion, such as when a person sits down.

V. Discussion & Conclusions

In this paper we have presented a simple algorithm that utilises a 3D modified five-point model for human pose estimation. We have presented the result to show that it has the potential for applications such as action recognition and gait recognition. Future work will address the development of improved decision making routines and the ability to include feedback from feature trackers for more robust tracking. The difficulty of estimating the human comes from the limited information contained in the 2D silhouettes. One way of improving the result is to compute the silhouettes to include shadow removal and body part identification. We also aim to adjust the flow of the algorithm by computing the 3D silhouettes then identifying the five points. The five-point model will also be extended to a full skeleton model in order to recover more sophisticated human motions.

This paper has illustrated the effectiveness of the modified five-point skeleton model for human pose recovery. It was shown that the technique was capable of correctly locating the 5 points with only a few exceptions. This was predominantly due to the heavy reliance on the silhouettes that were provided with the database. One major problem with these silhouettes are the shadows, which can be identified as local maximum and be labelled incorrectly as hands or legs depending on the viewing angle. The incorrect local maximum leads to incorrect 2D labelling. Triangulating the points from the 2D to 3D skeleton produces results that are far from accurate. This can be improved with shadow removal algorithms.

Another issue with the silhouette approach, similar to the occlusion problem, is that the silhouettes do not contain information visible in the scene. For example, when a person is standing with arms close to the body, the shoulders are likely to be detected as hands. Other improvements with calculating the maxima can be implemented with different approaches for calculating these distances. Instead of calculating the straight line distance from centroid to the edge, such as when traversing the line from the centroid to the candidate point, only one crossing is allowed. Even though the modified five-point skeleton model gives a somewhat realistic representation

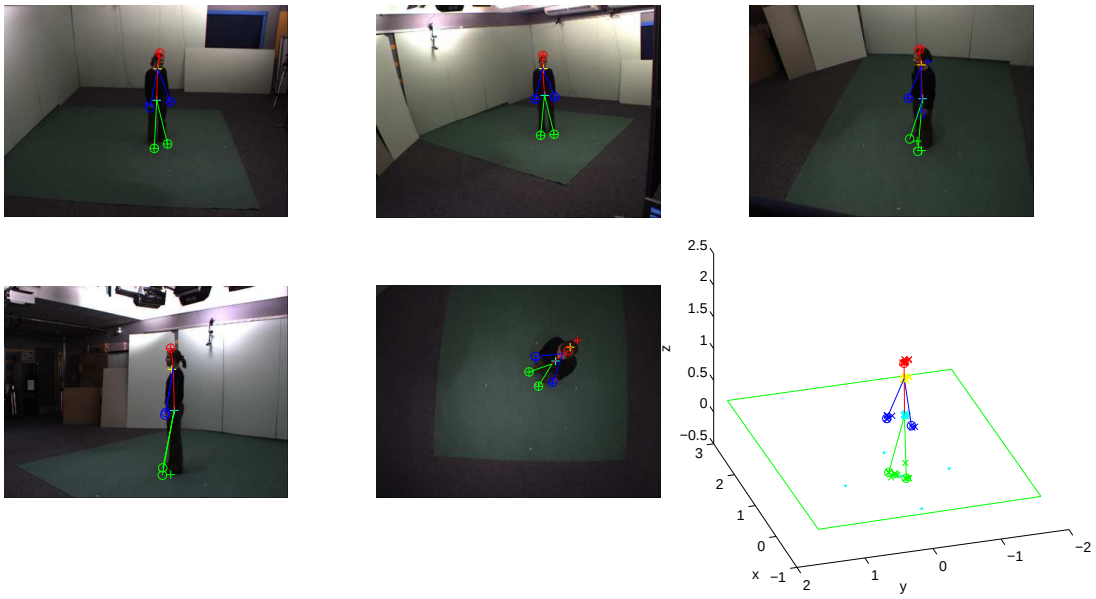

Fig. 4. Example frame showing the result of the algorithm

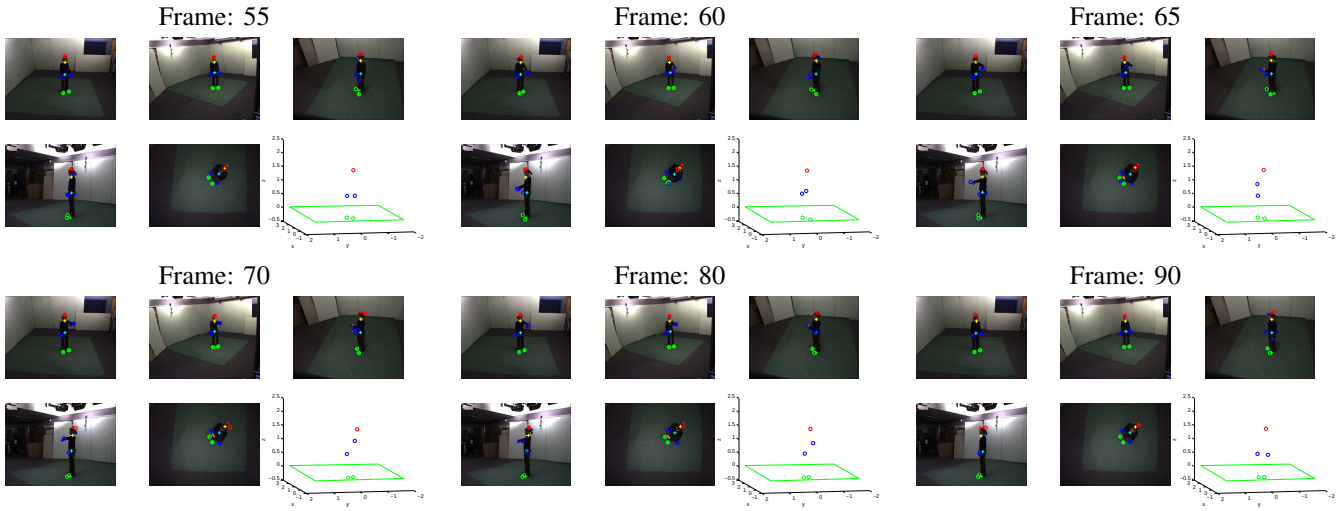

Fig. 5. 'Check watch' sequence

of the human pose, future work will concentrate on expansion to a full skeleton model to capture more detailed features from complex motions for more advanced human action recognition and gait recognition applications.

ACKNOWLEDGEMENTS

This project was supported by the Australian Government Department of the Prime Minister and Cabinet.

REFERENCES

[1] T. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.

[2] M. Nixon, J. Carter, D. Cunado, P. Huang, and S. Stevenage, "Automatic gait recognition," in *Biometrics*. Springer US, 2002, pp. 231–249.

[3] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, 1997, 0162-8828.

[4] R. Chellappa, A. K. Roy-Chowdhury, and S. K. Zhou, *Recognition of Humans and Their Activities Using Video*. Morgan & Claypool, 2005.

[5] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.

[6] H. Fujiyoshi and A. J. Lipton, "Real-time human motion analysis by image skeletonization," in *Fourth IEEE Workshop on Application of Computer Vision*, 1998, pp. 15–21.

[7] E. Yu and J. K. Aggarwal, "Detection of fence climbing from monocular video," in *18th International Conference on Pattern Recognition*, vol. 1, 2006, pp. 375–378.

[8] P. Peursum, H. Bui, S. Venkatesh, and G. West, "Human action recognition with an incomplete real-time pose skeleton," Curtin University of Technology, Tech. Rep. 2004/1, May 2004.

[9] N. Thome, D. Merad, and S. Miguet, "Human body part labeling and tracking using graph matching theory," in *IEEE International Conference*

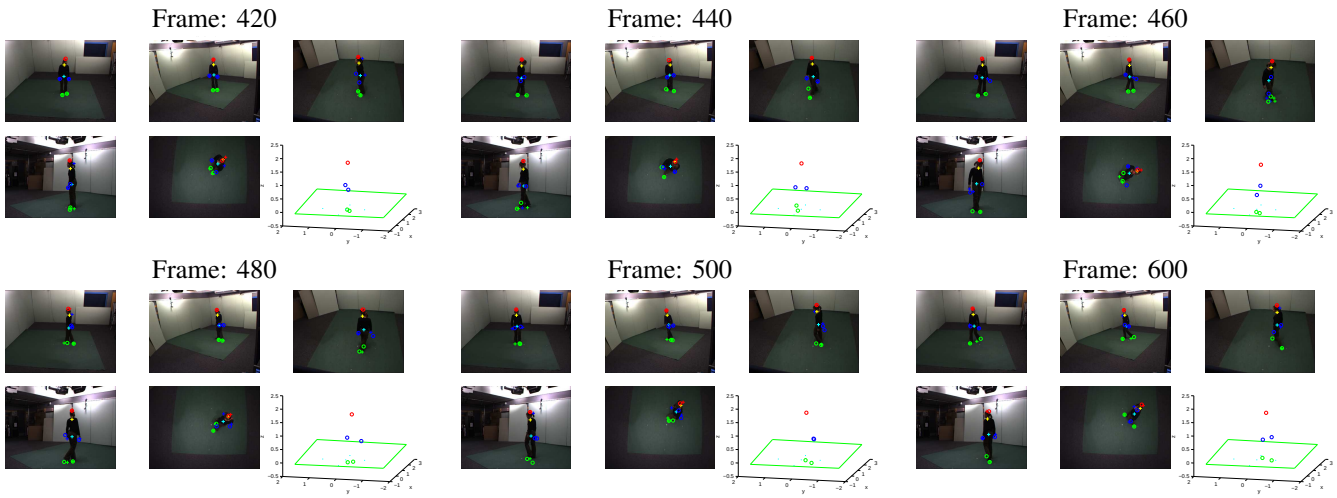

Fig. 6. 'Walk' sequence

on Video and Signal Based Surveillance, 2006.

[10] X. Ren, A. Berg, and J. Malik, "Recovering human body configurations using pairwise constraints between parts," in *Tenth IEEE International Conference on Computer Vision*, vol. 1, 2005, pp. 824–831.

[11] G. Mori, X. Ren, A. Efros, and J. Malik, "Recovering human body configurations: Combining segmentation and recognition," in *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2004, pp. 326–333.

[12] A. Agarwal and B. Triggs, "Recovering 3d human pose from monocular images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 44–58, 2006.

[13] C. Sminchisescu and B. Triggs, "Estimating articulated human motion with covariance scaled sampling," *International Journal of Robotics Research*, vol. 22, no. 6, pp. 371–391, 2003.

[14] D. M. Gavrila and L. S. Davis, "Towards 3-d model-based tracking and recognition of human movement," *International Workshop on Face and Gesture Recognition*, pp. 272–277, 1995.

[15] J. Lewis, "Fast normalized cross-correlation," in *Vision Interface*. Canadian Image Processing and Pattern Recognition Society, 1995, pp. 120–123.

[16] C. Tomasi and T. Kanade, "Detection and tracking of point features," Carnegie Mellon University, Tech. Rep. CMU-CS-91-132, April 1991.