Multiple Views Based Human Motion Tracking in Surveillance Videos

Bogdan Kwolek
Polish-Japanese Institute of Information Technology
ul. Koszykowa 86, 02-008 Warszawa, Poland
bkwolek@prz.edu.pl

Abstract

Most work on activity recognition focuses on 2D image properties, holistic spatiotemporal representations, or space-time shapes in image domain rather than with 3D pose in a body-centric or world frame. Such techniques rely on advanced pattern recognition algorithms and interpreting complex behavioral patterns. In this work we posit that it is possible to achieve 3D pose tracking using videos recorded in multi-camera surveillance systems. We show experimental results that were obtained on PETS 2009 datasets. The estimation of the 3D articulated motion is achieved using a modified particle swarm optimization.

1. Introduction

With millions of surveillance cameras monitoring city centers and streets, major meeting points like airports, underground and railway stations, automatic analysis of video content has become an important task. The fundamental problem in visual surveillance systems is detecting human presence, tracking human motion, analyzing human activities and signaling events of interest. A commonly-used strategy in vision-based surveillance systems is to detect people with bottom-up approaches such as background subtraction and color segmentation. The analysis and classification of human behavior inherently involves the estimation of body pose, understanding bodily motion, analysis of facial expressions, gait, etc. Multiple cameras with overlapping fields of view are usually utilized to disambiguate cluttered targets and to provide more confident reasoning about the events of interest. Such multi-view systems allow covering wide areas and handling the occurrence of occlusions by exploiting the different viewpoints.

Developing algorithms for activity recognition in surveillance videos that are both accurate and efficient in terms of computation overhead is challenging due to variability in shapes and articulation of human body, clutter, camera imperfections, lighting conditions and occlusion. Even when the scenario is simple, for instance there is only single unoccluded person and the illumination conditions are perfect, the action recognition is a difficult task, mainly because of variability and complexity of human actions.

Most algorithms for action and activity recognition were tested on video sequences acquired in controlled conditions or specific settings. A pioneering work on recognition of actions in less constrained conditions was done by Efros *et al*. [7]. The algorithm recognizes a set of simple actions of people whose images in the video are only 30 pixels tall and where the image quality is poor. As most current systems do not record videos at high resolution, dealing with low-resolution videos as well as decompressed images is important issue.

The majority of previous work on activity recognition has focused on 2D image properties, local interest points, holistic spatiotemporal representations, or space-time shapes in image domain rather than with 3D pose in a body-centric or world frame, *e.g*. [14]. Thus, most techniques rely on advanced pattern recognition algorithms and interpreting complex behavioral patterns that are generated when humans interact with others [20].

The experiments with moving light displays, which were done by Johansson [12] demonstrated that people are able to recognize human actions on the basis of motion of a small set of points on the human body. The discussed experiments stimulated discussion whether people recognize actions from 2D motion patterns, or they rely on 3D reconstructions from the motion of patterns. A discussion of motion-based techniques to activity recognition can be found in [3]. The authors argue that in action recognition, motion is more important modality than spatio-temporal representations. Gavrila [9] presents a survey focusing on the use of 2D or 3D models in action recognition. More recent survey on techniques related to activity recognition as well as motion capture and model initialization can be found in [16]. In general, the use of articulated models in action recognition simplifies occlusion handling. The model-based approaches are view independent and less dependent on the training data.

In recognition by reconstruction we can distinguish two

stages, namely motion capture stage that on the basis of an articulated 3D model estimates human motion and action recognition stage, which operates on joint trajectories. In motion-based techniques, the object trajectories are often used to infer object activities. As they are sensitive to translations, rotations and changes in scale many activity representations such as velocities, relative motions, spatio-temporal curvatures were proposed [3].

This work presents a 3D model based method for tracking human motion in surveillance videos. The articulated human body model is represented as a 3D kinematic tree consisting of 11 nodes. The tracking is accomplished with particle swarm optimization, with likelihood derived from background subtraction and edges. The algorithm estimates the 3D position of virtual markers at major joints of the body. In our approach we represent each motion as a pose trajectory, i.e., a vector consisting of the 3D joint positions at each time step. We report experimental results that were obtained on PETS 2009 datasets. The images were acquired by multiple cameras overlooking the same scene. The experimental results were obtained on decompressed images taken in real conditions. We demonstrate that the algorithm is able to cope with occlusions and imperfect synchronization between cameras. We demonstrate that owing to 3D articulated tracking there is no need for view-based models in behavior analysis.

The rest of the paper is organized as follows. In Section 2, we present background and the relevant work. Section 3 is devoted to particle swarm optimization. In Section 4, we outline our 3D articulated model and discuss the cost function. Finally, in Section 5, we present motion tracking results on PETS 2009 datasets. The paper concludes with a summary.

2. Background and related work

Our research was motivated by work of Sigal *et al*. [18] who demonstrated that on the basis of 3D articulated pose estimates it is possible to infer subtle physical attributes of human, like gender and weight, and even some aspects of mental state, *e.g*., happiness or sadness. As it was previously mentioned, the model-based approaches less dependent on the training data in comparison to methods that are based on holistic space-time features or space-time shapes. Moreover, by the use of 3D articulated tracking we do not need view-based models. One of the benefits of model-based trackers is that they permit a comprehensive exploration of the space of possible poses.

Several multi-camera approaches for articulated motion tracking were published recently. In their influential work, Gavrila and Davis [10] used four calibrated and widely spaced cameras. The model was projected into each of them under perspective. They defined a fitting cost by chamfer matching in filtered and background-subtracted edge images. The pose estimation was achieved by recursive search space decomposition, *i.e*., using a best-first search the best torso/head configuration is determined, afterwards the configuration of arms is estimated, etc. Kakadiaris and Metaxas [13] use three near-orthogonal views to perform 3D body tracking. An extended Kalman filtering is utilized for the model prediction. Furthermore, at each frame, in order to deal with occlusions a selection among 3 cameras takes place to choose views delivering the best information.

Deutscher *et al*. [6] demonstrated that probability density functions for joint angles are non-Gaussian. Even in case of the absence of cluttered background, the complex nature of the observation process during human motion capture causes the posterior density to be non-Gaussian. In [5], the motion tracking is achieved using three cameras and a particle filter based on simulated annealing. Compared to a classical particle filter, they reduced the number of samples by a factor of up to 10. In [2], instead of standard full projective geometry, scaled orthography is employed and in consequence the effects of changes in distance from the camera are compensated by changes in scale of the object. This approach seems to be appropriate for surveillance setups based on uncalibrated cameras or when objects of interest are far from the camera. In [17], a 3D model is used to achieve human motion capture in uncontrolled environments.

Particle filtering (PF) is one of the most important and common methods in human motion tracking. In the particle filter each sample represents some hypothesized body pose. However, the number of particles needed for a successful implementation of any PF algorithm grows exponentially with the dimension of the state. In contrast, particle swarm optimization (PSO) [15], which is a population-based searching technique, possesses better search efficiency by combining local search (by self experience) and global one (by neighboring experience). Particularly, a few simple rules result in high effectiveness in exploring the search space. Recently, PSO was proposed as an alternative of PF for full-body articulated motion tracking [21].

A survey on visual surveillance of object motion and behaviors is presented in [11]. Despite larger complexity, multiple camera setups exhibit several advantages consisting in covering wide areas, handling the occurrence of occlusions, reducing ambiguities in single camera's view. The main difficulty in multi-camera surveillance systems is establishing the relationship between multiple cameras and the corresponding object. In such systems, camera handoff is a fundamental step to obtain continuously tracked and consistently labeled trajectories of the objects of interest [4].

3. Object Tracking Using PSO

PSO is a population based algorithm introduced in [15] that utilizes a set of particles representing potential solu-

tions of the optimization task. Despite the simplicity of the individual particles, the swarm as a whole has a remarkable level of coherence and coordination. Each solution is represented as a series of coordinates in n-dimensional space. A number of particles are initialized randomly within the search space. Every particle flies in the solution space with a velocity adjusted dynamically according to its own experience and the experience of the whole swarm. Each particle has a very simple memory of its personal best solution so far, called $pbest$. The global best solution for each iteration is also determined and is termed $gbest$. On each iteration, every particle is moved a certain distance from its current location, influenced a random amount by the $pbest$ and $gbest$ values. The particles are evaluated according to a user defined fitness function $f()$. The velocity of each particle i is updated in accordance with the following equation:

$$v_i^{(j)} \leftarrow wv_i^{(j)} + c_1 r_1 (pbest_i^{(j)} - \omega_i^{(j)}) + c_2 r_2 (gbest^{(j)} - \omega_i^{(j)}) \quad (1)$$

where $v_i^{(j)}$ is the velocity in the $j-$th dimension of the $i-$th particle, c_1, c_2 denote the acceleration coefficients, r_1 and r_2 are uniquely generated random numbers in the interval $[0.0,\ 1.0]$, and w stands for an inertia weight. The inertia weight allows the balance of the exploration and exploitation abilities of the swarm as well as eliminates the need for velocity clamping.

The first part in (1) takes into account the previous velocity, which provides the necessary momentum for particles to fly across the search space. The second part is known as the cognitive component and represents the personal thinking of each particle. This component encourages the particles to fly toward their own best position $pbest$ found so far. The third part is known as the social component and represents the collaborative effect of the particles in finding the global optimum. This component pulls the particles toward the best position(s) found so far by their neighbors. The inertia part keeps particles to explore new areas while the cognitive and social parts try to keep them exploiting around the visited points.

The new position of a particle is calculated in the following manner:

$$x_i^{(j)} \leftarrow x_i^{(j)} + v_i^{(j)} \quad (2)$$

The local best position of each particle is updated as follows:

$$pbest_i \leftarrow \begin{cases} x_i, & \text{if } f(x_i) > f(pbest_i) \\ pbest_i, & \text{otherwise} \end{cases} \quad (3)$$

and the global best position $gbest$ is defined as:

$$gbest \leftarrow \arg\max_{pbest_i} \{f(pbest_i)\} \quad (4)$$

The value of velocity v_i should be restricted to the range $[-v_{max}, v_{max}]$ to prevent particles from moving out of the search range.

At the beginning of the optimization the PSO initializes randomly locations as well as the velocities of the particles. Then the algorithm selects $pbest$ and $gbest$ values. Afterwards, equations (1)-(4) are called until maximum iterations or minimum error criteria is attained.

In contrast to traditional optimization problems with stationary optima, tracking objects in image sequences requires the algorithm to find the optimum not once, but in every successive image. There are various approaches to dealing with moving objects, such as decaying the score of the best position after every frame. In consequence, such an operation results in forcing the swarm to continually search for a better location. In particular, it prevents the swarm from completely converging to a single point, allowing the swarm agents to be appropriately spaced in order to quickly reacquire a target in the next image. In PSO based tracking, at the beginning of each frame in the initialization stage, an initial position is assigned to each particle

$$\omega_{i,t} \leftarrow \mathcal{N}(gbest, \Sigma) \quad (5)$$

given the location $gbest$ that has been estimated in the previous frame $t-1$.

4. Tracking framework

The skeleton of the human body is modeled as a kinematic tree. The articulated 3D model is composed of a eleven segments with the limbs represented by truncated cones. The body model is built in a tree-like hierarchy starting with the pelvis as root body part. It comprises pelvis, torso/head, upper and lower arm and legs. The configuration of the model is defined by 20 joint angles plus global pose (26 degrees of freedom in total). The model's configuration is parameterized by position and orientation of the pelvis in the global coordinate system and the relative angles between the linked body segments. Together these parameters build a 26-dimensional configuration vector that specifies the pose of the model. From such a 3D representation of the human body, using a given parameters, any possible 2D view observation can be rendered.

In order to obtain a hypothesized 3D pose of the person of interest each truncated cone is projected into 2D image plane via perspective projection. In such manner we obtain an image containing the rendered model in a given configuration. Such features are then matched against observed person's features. In matching the model against the image cues the image edges are broadly used as they have good localization properties.

The fitness function consists of two components: $f(x) = w_1 f_1(x) + w_2 f_2(x)$, where w_i stands for weighting coefficients that were determined experimentally. The first term is silhouette-overlap term. It reflects the degree of overlap between the projected model and the extracted silhouette.

The silhouette-overlap degree is calculated through checking the overlap from the projected model to the extracted silhouette and from the silhouette to the rasterized image of the model. The silhouettes were delineated on the basis of the background subtraction [1]. The second term reflects the degree of overlap between the projected edges of the model and image gradients. Image gradients do not depend on background subtraction, but likewise are sensitive to object properties, textures, and lighting, etc. However, they might be easily confused with static object in the background with strong gradients. Therefore, in the second term we employ background-subtracted edge images. Additionally, the second term is calculated with the support of edge-proximity, *i.e*. distance map. A distance map is essentially a grey level image where the pixel intensity is determined by its distance from the nearest edge. In our approach we employ chessboard distance and limit the number of iterations on the chain propagation to 3. Figure 1 depicts the images that are utilized in the calculation of the cost function.

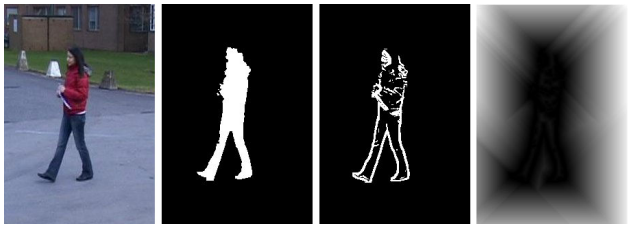

Figure 1. Input image, background subtraction, thresholded gradient magnitude, distance map.

The estimation of the pose is achieved by a modified particle swarm optimization. Our approach is motivated by [10], where the configuration space was constrained using a hierarchical search. In the discussed approach, called search space decomposition, a part of the articulated model is localized independently in advance, and then its location is used to constrain the search for the remaining limbs. On the basis of color cues the torso is localized first and then it is used to confine the search for the limbs. However, in realistic scenarios, among others due to occlusions, it is not easy to localize the torso and to extract reliably such a good starting guess for the search. Therefore in our approach we first localize the torso using the whole model and reduced number of particles. Afterwards, given the configuration of the legs in previous frame, we carry out rediversification of the particles in the part of the vector state that describes the pose of the legs. This means, that given the location of the torso that has been determined in advance, we generate several hypothesized leg configurations. Then we perform optimization using part of the state vector that describes the pose of the legs. At this stage, in the objective function we employ only legs. The pose of the hands is determined in a similar manner.

5. Experiments

The method has been evaluated on the PETS 2009 datasets [8]. The PETS sequences were recorded at 7 frames per second. The images of size 720×576 were taken from different positions using different cameras. All frames are compressed as JPEG image sequences. The cameras were calibrated using Tsai model [19] and geometric patterns on the ground. Although every effort has been made to ensure the synchronization of frames from different views, there exist slight delays and even frame drops. The experiments were conducted on Dataset S2, which addresses people tracking. Figure 2 shows representative frames from the views 5, 6 and 8, which were utilized in our tests. The datasets from views 5 and 6 were filmed using Sony DCR-PC1000E (ffmpeg de-interlaced), whereas the dataset from view 8 was filmed using Canon MV-1 (progressive scan).

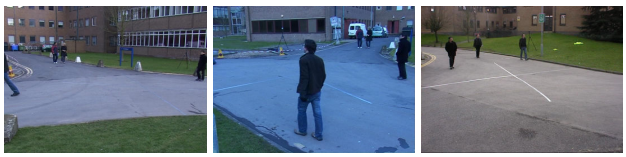

Figure 2. Representative frames from the sequence S2 and views 5, 6 and 8.

Figure 3 depicts foreground images, which were extracted from the images shown at Fig. 2. The foreground images were used to compute the silhouette-overlap term in the cost function. The images were dilated in order to extract binary masks of the persons. They were then used to extract background-subtracted edge images.

Figure 3. Representative binary images from the sequence S2 and views 5, 6 and 8.

At Fig. 4 we can see some tracking results that were obtained in frames 476-499 using images from views 5, 6 and 8. The tracking was initialized manually in frame #475. As we can observe, the motion tracker is able to cope with temporal occlusions in a camera view, see frame #476.

Images in the second row of Fig. 5 illustrate the estimated pose in consecutive frames. The motion estimation has been achieved using images from views 5, 6 and 8. As we can observe, the lateral walk pose is estimated quite reliably using the mentioned camera views. In the third row we can observe the location of some virtual markers. Since the human body consists of several rigidly moving body limbs,

Figure 4. Motion tracking. Frames #476, 481, 487, 499.

the pose can be seen as a set of parameters that describe the actual location of these body parts. People easily recognize the human motions if only the positions of the major skeletal joints (i.e., the elbows, shoulders, ankle, etc.) are visible, e.g. as white moving dots on a black background, as depicted graphically at images in third row of Fig. 5.

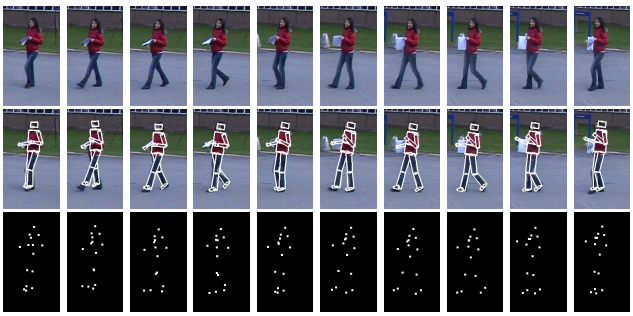

Figure 5. Motion tracking in frames 475-484. Croped input images (1st row), estimated pose (2nd row), position of virtual markers (3rd row).

The algorithm delivers joint angles in body-centric coordinate frame. It generates also the trajectories of the virtual markers, see Fig. 6. Of course there are other possible motion features [3].

Figure 6. Trajectories of virtual markers.

Figure 7 illustrates motion tracking, which was initialized in frame #600. The motion of the person has been successfully tracked until frame #643. As we can observe, the algorithm can deal with temporal occlusions, see frame #608 in view 6, as well as frames #613 and #618 in view 8.

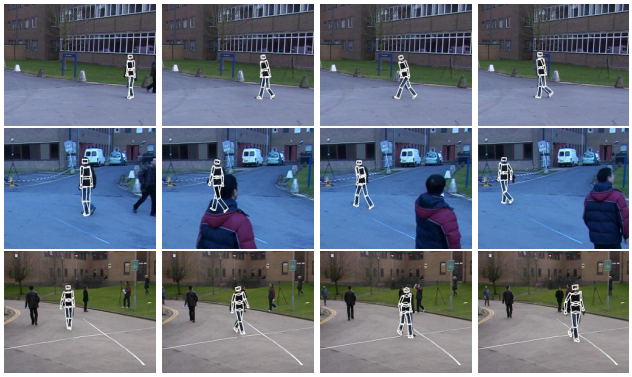

Figure 7. Motion tracking. Frames #601, 608, 613, 618.

At Fig. 8 are demonstrated some results of motion tracking, which has been initialized in frame #231. The motion of the person has successfully been tracked until frame #266 despite partial occlusion in frames 237-241 from view 6 and considerable occlusion in frames 254-264 from view 8.

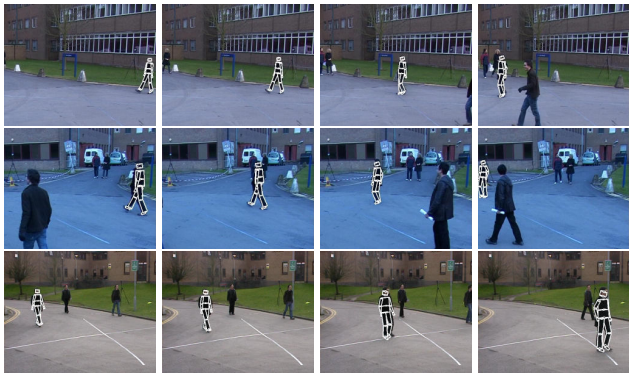

Figure 8. Tracking result in frames #232, 241, 251, 267.

Figure 9 depicts the 3D locations and velocities of left and right ankle and knee. The larger the cone is, the larger is the velocity of the virtual marker.

The algorithm has been implemented in C/C++. The experiments were conducted on desktop PC with 4 GB RAM, Intel Core i5, 2.8 GHz. The parallelization of the code was done using OpenMP directives and the parallel computations were realized on multi-core (4-core) CPUs. The algorithm operates at $\sim$0.65 fps. The above presented results were achieved by PSO in 20 iterations in each phase. During estimation of the whole pose we used 100 particles, whereas while estimation of the pose of each limb pair we utilized 50 particles.

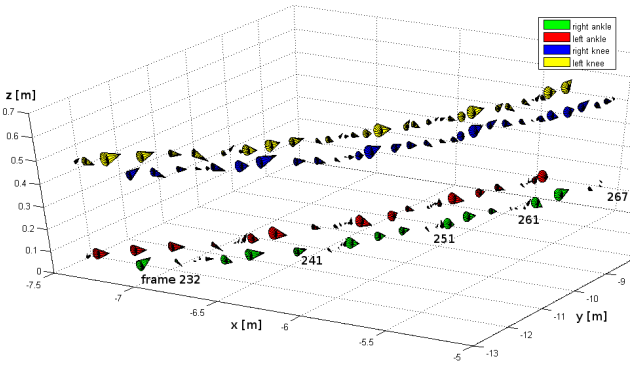

Figure 9. 3D trajectories and velocities of left/right ankle and knee.

6. Conclusions

In this paper, we have shown that a successful articulated motion tracking can be done on surveillance videos. The results showed that even in case of slight delays between frames from different views the tracking is still possible. Moreover, the results demonstrated that successful tracking is possible even in case of temporal occlusion in one of the camera views. We demonstrated that 3D human poses can contribute toward view-invariant action recognition. While video-based motion estimates are noisy, they can support action recognition in surveillance videos.

Acknowledgement

This paper has been supported by the research project OR00002111: "Video surveillance systems for person and behavior identification and threat detection, using biometrics and inference of 3D human pose from video."

References

[1] D. Arsic, A. Lyutskanov, G. Rigoll, and B. Kwolek. Multi camera person tracking applying a graph-cuts based foreground segmentation in a homography framework. In *IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance*, pages 30–37. IEEE Press, 2009.

[2] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proc. IEEE Int. Conf. on Comp. Vision and Pattern Rec.*, pages 8–15. IEEE Computer Society, 1998.

[3] C. Cdras and M. Shah. Motion-based recognition a survey. *Image and Vision Computing*, 13:129–155, 1995.

[4] C.-H. Chen, Y. Yao, D. Page, B. Abidi, A. Koschan, and M. Abidi. Camera handoff with adaptive resource management for multi-camera multi-object tracking. *Image and Vision Computing*, 28(6):851 – 864, 2010.

[5] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *IEEE Int. Conf. on Computer Vision and Pattern Rec.*, pages 126–133, 2000.

[6] J. Deutscher, B. North, B. Bascle, and A. Blake. Tracking through singularities and discontinuities by random sampling. In *Proc. Int. Conf. on Computer Vision - Vol. 2*, pages 1144–1149. IEEE Computer Society, 1999.

[7] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proc. of the IEEE Int. Conf. on Computer Vision - Vol. 2*, ICCV '03, pages 726–733, Washington, DC, USA, 2003. IEEE Computer Society.

[8] J. Ferryman and A. Shahrokni. Pets2009: Dataset and challenge. In *IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance*, pages 1–6, Dec. 2009.

[9] D. M. Gavrila. The visual analysis of human movement: a survey. *Comput. Vis. Image Underst.*, 73:82–98, 1999.

[10] D. M. Gavrila and L. S. Davis. 3-D model-based tracking of humans in action: a multi-view approach. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Rec.*, pages 73–80, Washington, DC, USA, 1996. IEEE Computer Society.

[11] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems Man and Cybernetics, Part C Applications and Reviews*, 34:334–352, 2004.

[12] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perceptron & Psychophysics*, 14:201–211, 1973.

[13] I. A. Kakadiaris and D. Metaxas. Model-based estimation of 3d human motion with occlusion based on active multi-viewpoint selection. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 81–87, Washington, DC, USA, 1996. IEEE Computer Society.

[14] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *IEEE Int. Conf. on Computer Vision*, page 18, 2007.

[15] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proc. of IEEE Int. Conf. on Neural Networks*, pages 1942–1948. IEEE Press, Piscataway, NJ, 1995.

[16] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.*, 104:90–126, November 2006.

[17] M. Shaheen, J. Gall, R. Strzodka, and H.-P. Seidel. A comparison of 3d model-based tracking approaches for human motion capture in uncontrolled environments. In *IEEE Workshop on Appl. of Comp. Vision*, pages 1–8, December 2009.

[18] L. Sigal, D. J. Fleet, N. F. Troje, and M. Livne. Human attributes from 3d pose tracking. In *Proc. of the 11th European Conf. on Computer vision: Part III*, ECCV'10, pages 243–257, Berlin, Heidelberg, 2010. Springer-Verlag.

[19] R. Tsai. An efficient and accurate camera calibration technique for 3d machine vision. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 323–344, 1986.

[20] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.*, 115:224–241, February 2011.

[21] X. Zhang, W. Hu, X. Wang, Y. Kong, N. Xie, H. Wang, H. Ling, and S. Maybank. A swarm intelligence based searching strategy for articulated 3D human body tracking. In *IEEE Workshop on 3D Information Extraction for Video Analysis and Mining in conjuction with CVPR*, pages 45–50. IEEE Press, Piscataway, NJ, 2010.