Available online at www.sciencedirect.com
ScienceDirect
Computer Vision and Image Understanding 108 (2007) 4–18
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

Vision-based human motion analysis: An overview

Ronald Poppe

Human Media Interaction Group, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, P.O. Box 217, 7500 AE, Enschede, The Netherlands

Received 20 September 2005; accepted 13 October 2006
Available online 25 January 2007
Communicated by Mathias Kolsch

Abstract

Markerless vision-based human motion analysis has the potential to provide an inexpensive, non-obtrusive solution for the estimation of body poses. The significant research effort in this domain has been motivated by the fact that many application areas, including surveillance, Human–Computer Interaction and automatic annotation, will benefit from a robust solution. In this paper, we discuss the characteristics of human motion analysis. We divide the analysis into a modeling and an estimation phase. Modeling is the construction of the likelihood function, estimation is concerned with finding the most likely pose given the likelihood surface. We discuss model-free approaches separately. This taxonomy allows us to highlight trends in the domain and to point out limitations of the current state of the art.

Keywords: Human motion analysis; Pose estimation; Computer vision

1. Introduction

Human body pose estimation, or pose estimation in short, is the process in which the configuration of body parts is estimated from sensor input. When poses are estimated over time, the term human motion analysis is used. Traditionally, motion capture systems require that (electromagnetic) markers are attached to the body. These systems have two major drawbacks: they are obtrusive and expensive. Many applications, especially in surveillance and Human–Computer Interaction (HCI), would benefit from a solution that is markerless. Vision-based motion capture systems attempt to provide such a solution, using cameras as sensors. Over the last two decades, this topic has received much interest, and it continues to be an active research domain. In this overview, we summarize the characteristics of and challenges presented by markerless vision-based human motion analysis. The literature is discussed, with a focus on recent work. However, we do not intend to give complete coverage to all work.

1.1. Scope of this overview

Human motion analysis is a broad concept. In theory, as many details as the human body can exhibit could be estimated. This includes facial movement, movement of the fingers and changes in skin surface as a result of muscle tightening. In this overview, pose estimation is limited to large body parts (trunk, head, limbs). Note that, in human motion analysis, we are only interested in the configurations of the body parts over time and not interpretations of the movement. This means that pose recognition, which is classifying the pose to one of a limited number of classes, and gesture recognition, which is interpreting the movement over time, are not discussed in this overview. For some applications, the positioning of individual body parts is not important. The entire body is tracked as a single object, which is termed human tracking or detection. This is often a preprocessing step for human motion analysis, and we will not discuss the topic in detail in this overview. Surveys of literature on related fields can be found in [78,25] (gesture recognition), and [125] (face recognition).

E-mail address: poppe@ewi.utwente.nl

1077-3142/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2006.10.016

In the remainder of this section, we summarize past surveys and taxonomies, and describe the taxonomy that is used throughout this overview.

1.2. Surveys and taxonomies

Within the domain of human motion analysis, several surveys have been written, each with a specific focus and taxonomy. Gavrila [27] divides research into 2D and 3D approaches. 2D approaches are further subdivided into approaches with or without the explicit use of shape models. Aggarwal and Cai [4] use a taxonomy with three categories: body structure analysis, tracking and recognition. Body structure analysis is essentially pose estimation and is split up into model-based and model-free, depending upon whether *a priori* information about the object shape is employed. A taxonomy for tracking is divided into single and multiple perspectives. Moeslund and Granum [63,64] use a taxonomy based on subsequent phases in the pose estimation process: initialization, tracking, pose estimation and recognition. Wang et al. [121] use a taxonomy similar to [4]: human detection, human tracking and human behavior understanding. Tracking is subdivided into model-based, region-based, active contour-based and feature-based. Wang and Singh [120] identify two phases in the process of computational analysis of human movement: tracking and motion analysis. Tracking is discussed for hands, head and full bodies.

Currently, we see some new directions of research such as combining top–down and bottom–up models, particle filtering algorithms for tracking, and model-free approaches. We feel that many of these trends cannot be discussed appropriately within the taxonomies mentioned above. We observe that studies can be divided into two main classes: model-based (or generative) and model-free (or discriminative) approaches. Model-based approaches employ an *a priori* human body. The pose estimation process consists of modeling and estimation [100]. Modeling is the construction of the likelihood function, taking into account the camera model, the image descriptors, human body model and matching function, and (physical) constraints. We discuss the modeling process in detail in Section 2. Estimation is concerned with finding the most likely pose given the likelihood surface. The estimation process is discussed in Section 3. Model-free approaches do not assume an *a priori* human body model but implicitly model variations in pose configuration, body shape, camera viewpoint and appearance. Due to their different nature in both modeling and estimation, we discuss them separately in Section 4. We conclude with a discussion of open challenges and promising directions of research.

2. Modeling

The goal of the modeling phase is to construct the function that gives the likelihood of the image, given a set of parameters. These parameters include body configuration parameters, body shape and appearance parameters and camera viewpoint. Some of these parameters are assumed to be known in advance, for example a fixed camera viewpoint, or known body part lengths. Estimating a smaller number of parameters makes the problem more tractable but also poses limitations on the visual input that can be appropriately analyzed. Note that the relation between pose and observation is multivalued, in both directions. Due to the variations between people in shape and appearance, and a different camera viewpoint and environment, the same pose can have many different observations. Also, different poses can result in the same observation. Since the observation is a projection (or combination of projections when multiple cameras are deployed) of the real world, information is lost. When only a single camera is used, depth ambiguities can occur. Also, because the visual resolution of the observations is limited, small changes in pose can go unnoticed.

Model-based approaches use a human body model, which includes the kinematic structure and the body dimensions. In addition, a function that describes how the human body appears in the image domain, given the model's parameters, is used. Human body models are described in Section 2.1.

Instead of using the original visual input, the image is often described in terms of edges, color regions or silhouettes. A matching function between visual input and the generated appearance of the human body model is needed to evaluate how well the model instantiation explains the visual input. Image descriptors and matching functions are described in Section 2.2. Other factors that influence the construction of the likelihood function are the camera parameters (Section 2.3) and environment settings (Section 2.4).

2.1. Human body models

Human body models describe both the kinematic properties of the body (the skeleton), as the shape and appearance (the flesh and skin). We discuss both below.

2.1.1. Kinematic models

Most of the models describe the human body as a kinematic tree, consisting of segments that are linked by joints. Every joint contains a number of degrees of freedom (DOF), indicating in how many directions the joint can move. All DOF in the body model together form the pose representation. These models can be described in either 2D or 3D.

2D models are suitable for motion parallel to the image plane and are sometimes used for gait analysis. Ju et al. [44], Haritaoglu et al. [33] and Howe et al. [38] use a so-called Cardboard model in which the limbs are modeled as planar patches. Each segment has seven parameters that allow it to rotate and scale according to the 3D motion. Navaratnam et al. [70] take a similar approach but model some parameters implicitly. In [40], an extra patch width

parameter was added to account for scaling during in-plane motion. In [16,1], the human body is described by a 2D scaled prismatic model [68]. These models have fewer parameters and enforce 2D constraints on figure motion that are consistent with an underlying 3D kinematic model. But despite their success in capturing fronto-parallel human movement, the inability to encode joint angle limits and self-intersection constraints renders 2D models unsuitable for tracking more complex movement.

3D models most often model segments as rigid, and allow a maximum of three (orthogonal) rotations per joint. For each of the rotations individually, kinematic constraints can be imposed. Instead of segments that are linked with zero-displacement, Kakadiaris and Metaxas [46] model the connection by constraints on the limb ends. In a similar fashion, Sigal et al. [99] model the relationships between body parts as conditional probability distributions. Bregler et al. [13] introduce a twist motion model and exponential maps which simplify the relation between image motion and model motion. The kinematic DOF can be recovered robustly by solving simple linear systems under scaled orthogonal projection.

The number of DOF that are recovered varies between studies. In some studies, a mere 10 DOF are recovered in the upper body. Other studies estimate full-body poses with no less than 50 DOF [3,5]. But even for a model with a limited number of DOF and a coarse resolution in (discrete) parameter space, the number of possible poses is very high. Applying kinematic constraints is an effective way of pruning the pose space by eliminating infeasible poses. Typical constraints are joint angle limits [118,21] and limits on angular velocity and acceleration [124]. The fact that human body parts are non-penetrable also introduces constraints [105].

2.1.2. Shape models

Apart from the kinematic structure, the human shape is also modeled. Segments in 2D models are described as rectangular or trapezoid-shaped patches (see Fig. 1(a)). In 3D models segments are either volumetric or surface-based. Volumetric shapes depend on only a few parameters. Commonly used volumetric models are spheres [74], cylinders [34,87,93] or tapered super-quadrics [19,28,47] (see Fig. 1(b)). Instead of modeling each segment as a separate rigid shape [15], surface-based models often employ a single surface for the entire human body (see Fig. 1(c)). These models typically consist of a mesh of polygons that is deformed by changes to the underlying kinematic structure [5,45,9]. Plänkers and Fua [79] use a more complex body shape model, consisting of three layers: kinematic model, metaballs (soft objects) and a polygonal skin surface.

The parameters of the shape model, such as shape lengths and widths, are sometimes assumed fixed. However, due to the large variability among people, this will lead to inaccurate pose estimations. Alternatively, these parameters can be recovered in an initialization step, where the observed person is to adopt a specified pose [15,6]. While this approach works well for many applications, it restricts use in surveillance or automatic annotation systems. Online adjustment of these parameters is possible by relying on statistical priors [30] or specific key poses [18,8]. Cheung et al. [17] and Mikić et al. [61] use a number of cameras and recover segment shape and joint positions by looking at motion of individual points. Krahnstöver et al. [49] report similar work for the upper body using a single camera but only seem to support movement parallel to the image plane.

The likeliness of the model instantiation given the image can be calculated when functions are available that describe how the model instantiation appears in the image domain and calculate the distance between given image and synthesized model. We describe model appearance in the image domain, and the matching functions, in Section 2.2.

2.2. Image descriptors

The appearance of people in images varies due to different clothing and lighting conditions. Since we focus on the recovery of the kinematic configuration of a person, we would like to generalize over these kinds of variation. Part of this generalization can be handled in the image domain by extracting image descriptors rather than taking the original image. From a synthesis point of view, this means that we do not need complete knowledge about how a model instantiation appears in the image domain. Often used image descriptors include silhouettes, edges, 3D reconstructions, motion and color. We describe these next.

2.2.1. Silhouettes and contours

Silhouettes and contours (silhouette outlines) can be extracted relatively robustly from images when backgrounds are reasonably static. In older studies, backgrounds were often assumed to be different in appearance from the person. This eliminates the need to estimate environment parameters. Silhouettes are insensitive to variations in surface such as color and texture, and encode a great deal of information to help recover 3D poses [3]. However, performance is limited due to artifacts such as shadows and noisy background segmentation, and it is often difficult or impossible to recover certain DOF due to the lack of depth information (see Fig. 2). A matching function is often based on area overlap. In model-free approaches, silhouettes are encoded using central moments [11] or Hu moments [89]. Contours can be encoded using a combination of turning angle metric and Chamfer distance [35] or shape contexts [7], and can be compared based on deformation cost [66].

2.2.2. Edges

Edges appear in the image when there is a substantial difference in intensity at different sides of the image location. Edges can be extracted robustly and at low cost. They are, to some extent, invariant to lighting conditions, but are unsuitable when dealing with cluttered backgrounds or tex-

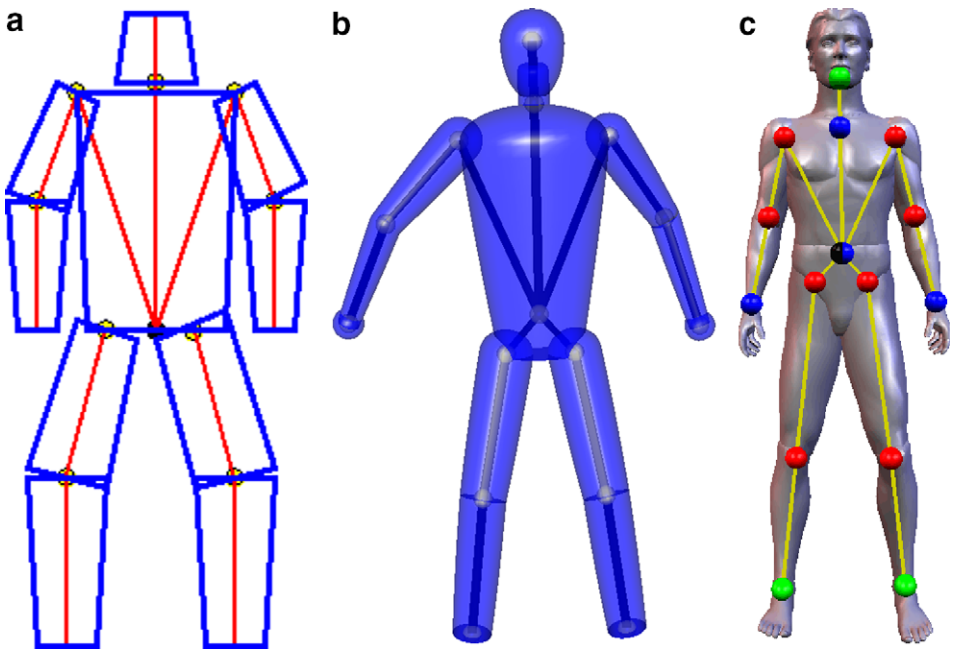

Fig. 1. Human shape models with kinematic model. (a) 2D model (reprinted from [40], © IEEE 2002); (b) 3D volumetric model consisting of superquadrics (reprinted from [47], © Elsevier, 2006); (c) 3D surface model (reprinted from [15], © ACM, Inc., 2003).

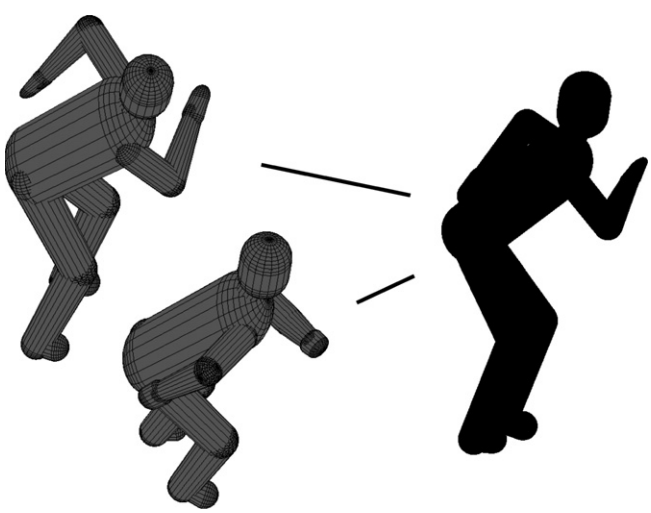

Fig. 2. Depth ambiguities when using monocular silhouettes [35] (© IEEE, 2004).

tured clothing. Therefore, edges are usually located within an extracted silhouette [46,118,87] or within a projection of a human model [23]. Matching functions take into account the normalized distance between model's synthesized edges and the closest edge found in the image. Rohr [87] uses edge lines instead of edges to partially eliminate silhouette noise. A distance measure based on difference in line segment length, center position and angle is applied.

2.2.3. 3D reconstructions

Edges and silhouettes lack depth information, at least when only a single camera is used. This also makes it hard to detect self-occlusions. When multiple cameras are used, a 3D reconstruction can be created from silhouettes that are extracted in each view individually. Two common techniques are volume intersection [9] or a voxel-based approach [17,61].

Another way of obtaining depth information is by using stereometry. Corresponding points are sought in views of calibrated camera pairs. Using triangulation, the depths of the points are calculated. This approach has been taken by Plänkers and Fua [79] and Haritaoglu et al. [33]. Stereo is also used by Jojic et al. [43], with the optional aid of projected light patterns. Matching functions are based volume overlap or mean closest point distance.

2.2.4. Color and texture

Modeling the human body based on color or texture is inspired by the observation that the appearance of individual body parts remains substantially unchanged, although the body may exhibit very different poses. The appearance of individual body parts can be described using Gaussian color distributions [123] or color histograms [81]. Roberts et al. [85] propose a 3D appearance model to overcome the problems with changing appearance due to clothing, illumination and rotations. They model body parts with truncated cylinders, with surface patches described by a multi-modal color distribution. The appearance model is constructed on-line from monocular image streams. Barrón and Kakadiaris [6] minimize the sum of pixel-wise intensity differences between the image and synthesized model. Skin color can be a good cue for finding head and hands. In [53], additional clothing parameters are used to model sleeve, hem and sock lengths.

2.2.5. Motion

Motion can be measured by taking the difference between two consecutive frames. The brightness of the pixels that are part of the person in the image are assumed to be constant. The pixel displacement in the image is termed optical flow and is used by Bregler et al. [13] and Ju et al. [44]. Sminchisescu and Triggs [105] use optical flow to construct an outlier map that is used to weight the importance of edges.

2.2.6. Combination of descriptors

A likelihood function that takes into account a combination of descriptors proves to be more robust. Silhouette information can be combined with edges [21], optical flow [36] or color [17]. In [92], edges, ridges and motion are used. Filter responses for these image cues are learned from training data. Ramanan and Forsyth [81] use edges and appearance cues. Care must be taken in constructing the likelihood function, especially when multiple image descriptors are used. Not unusually, a body part configuration that results in a low cost for one image descriptor, will also result in a low cost for a second one. When the likelihood function simply multiplies the cost function for each image descriptor, this may lead to sharp peaks in the likelihood surface. This results in less effective estimation.

2.3. Camera considerations

Regarding the number of cameras that is used, monocular work [38,3,105,93] is appealing since for many applications only a single camera is available. When only a single view is used, self-occlusions and depth ambiguities can occur. Sminchisescu and Triggs [105] estimate that roughly one third of all DOF are almost unobservable. These are mainly motions in depth but also rotations of near-cylindrical limbs about their axes. These limitations can be alleviated by using multiple cameras. In general, there are two main approaches. One is to search for features in each camera image separately and in a later stage combine the information to resolve ambiguities [19,28,90,83]. The second approach is to combine the information as early as possible into a 3D reconstruction, as we described before. When multiple cameras are used, calibration is an important requirement. Instead of combining the views, Kakadiaris and Metaxas [46] use active viewpoint selection to determine which cameras are suitable for estimation.

Most studies assume a scaled orthographic projection which limits their use to distant observations, where perspective effects are small. Rogez et al. [86] remove the perspective effect in a preprocessing step.

2.4. Environment considerations

Most of the approaches described in this overview can handle only a single person at a time. Pose estimation of more than one person at the same time is difficult because of occlusions and possible interactions between the persons. However, Mittal et al. [62] were able to extract silhouettes of all persons in the scene using the M_2Tracker. A setup with five cameras provides the input for their method. The W^4S system [33] is able to track multiple persons and estimate their poses in outdoor scenes using stereo image pairs and appearance cues.

The results that are obtained are largely influenced by the complexity of the environment. Outdoor scenes are much more challenging due to the dynamic background and lighting conditions. In most work, the persons are visible without occlusion by other objects. It remains a challenge to recover poses of people under significant occlusion.

3. Estimation

The estimation process is concerned with finding the set of pose parameters that minimizes the error between observations on the one hand, and on the other the projection of the human body model (model-based), projection function (learning-based) or example set (example-based). We can identify two classes of estimation: top–down and bottom–up. Top–down approaches match a projection of the human body with the image observation. Instead, in bottom–up approaches individual body parts are found and then assembled into a human body. Recent work combines these two classes. We discuss both classes and their combination in Section 3.1.

The likelihood function often has many local maxima [106]. In this section, we will assume that instead of a likelihood function, a cost function has been constructed. Therefore, we search for minima instead of maxima. Given the high dimensionality of the search space, this search must be efficient. The speed of the pose recovery depends largely on the speed of the estimation strategy. Some approaches report estimation times of several minutes per frame, other approaches can estimate human motion in real time [23].

Many methods are single-hypothesis approaches. Recent studies maintain multiple hypotheses. This reduces the probability of getting stuck at a local minimum. We discuss single and multiple hypothesis tracking, and batch methods, in Section 3.2.

Estimation of poses over time can be made more stable by assuming a motion model. Usually, these models are specific for a given activity. In Section 3.4, both explicit and implicit motion models are discussed.

3.1. Top–down and bottom–up estimation

There are two main approaches for model-based estimation: top–down and bottom–up. Recent work combines these approaches to benefit from the advantages of both.

3.1.1. Top–down estimation

Top–down approaches match a projection of the human body with the image observation. This is termed an analy-

sis-by-synthesis approach. A local search is often performed around an initial pose estimate [28,13,6]. A brute-force local search is computationally expensive due to the high dimensionality of the pose space. Therefore, the *a posteriori* pose estimate is often found by applying gradient descent on the cost surface [118]. The search can also be performed in the image domain. Delamarre and Faugeras [19] use forces between extracted silhouettes and the projected model to refine the pose estimation. Alternatively, sampling-based approaches are taken. We discuss these in the next section.

One drawback of top–down estimation is the fact that (manual) initialization in the first frame of a sequence is needed since the initial estimate is often obtained from the estimate in the previous frame. Another drawback is the computational cost of forward rendering the human body model and calculating the distance between the rendered model and the image observation.

Gavrila and Davis [28] take a top–down approach with search-space decomposition. Poses are estimated in a hierarchical coarse-to-fine strategy, estimating the torso and head first and then working down the limbs. The initial pose prediction is based on constant joint angle acceleration. An analysis-by-synthesis approach is applied in a discrete fashion, resulting in a limited number of possible solutions per joint.

Top–down estimation often causes problems with (self)occlusions. Moreover, errors are propagated through the kinematic chain. An inaccurate estimation for the torso/head part causes errors in estimating the orientation of body parts lower in the kinematic chain. To overcome this problem, Drummond and Cipolla [23] introduce constraints between linked body parts in the kinematic chain. This allows lower parts to effect parts higher in the chain. A pose is described by the rigid displacement for each body part. This yields an over-parameterized system which is solved in a weighted least-squares framework.

3.1.2. Bottom–up estimation

Bottom–up approaches are characterized by finding body parts and then assembling these into a human body. The body parts are usually described by 2D templates. Often, these templates produce many false positives, as there are often many limb-like regions in an image. Another drawback is the need for part detectors for most body parts, since missing information is likely to result in a less accurate pose estimate.

The assembling process takes into account physical constraints such as body part proximity. Temporal constraints can be used to cope with occlusions. Bottom–up approaches have the advantage that no manual initialization is needed and can be used as an initialization for top–down approaches.

Mori et al. [67] first perform image segmentation based on contour, shape and appearance cues. The segments are classified by body part locators for half-limbs and torso that are trained on image cues. From this partial configuration, the missing body parts are found. Global constraints, including body part proximity, relative widths and lengths and symmetry in color are enforced to prune the search space. A very similar approach has been taken by Ren et al. [84], who search for pairwise edges as segment boundaries. Ramanan [80] improves the deformable model iteratively, but does not perform explicit segmentation. In the first iteration, only edges are used to locate possible body parts. A rough region-based model for each body part and the background is then build from these locations. New locations are found using this model and the process is repeated.

In [26] body parts are modeled using 2D appearance models. They use the concept of pictorial structures to model the coherence between body parts. An efficient dynamic programming algorithm is used to find an optimal solution in the tree of body configurations. Trees are extended with correlations between body parts in [50]. For walking, correlations between upper arm and leg swings are used, resulting in more robust pose estimations. Ronfard et al. [88] use the pictorial structures concept but replace the body part detectors by more complex ones that learn appearance models using Support Vector Machines. Ramanan and Forsyth [81] use simple appearance-based part detectors, aided by parallel lines. Motion tracking is reduced to the problem of inference in a dynamic Bayes net. Evaluation on outdoor sequences shows automatic initialization and recovery but tracking occasionally fails, especially for in-plane motion. Ioffe and Forsyth [41] also take a 2D approach where the appearance of individual body parts is modeled. Inference is used on a mixture of trees, to avoid the time consuming evaluation of each group of candidate primitives. Song et al. [107] use a similar technique involving feature points and inference on a tree model.

Sigal et al. [99] describe the human body as a graphical model where each node represents a parameterized body part (see Fig. 3(a)). The spatial constraints between body parts are modeled as arcs. Each node in the graph has an associated image likelihood function that models the probability of observing image measurements conditioned on the position and orientation of the part. Pose estimation is simply inference in the graphical model. In [95,32], temporal constraints are also taken into account, resulting in a tracking framework. If individual part locators are used, there is the risk that the estimated pose does not explain the image very well. Sigal and Black [97] introduced occlusion-sensitive image likelihoods, which introduces loops in the graphical model. Recently, they focussed on obtaining 3D poses from these 2D pose descriptions [98].

Ramanan and Sminchisescu [82] train models that maximize the likelihood for joint localization of all body parts, rather than learning individual part locators. Their training algorithm learns the parameters of a Conditional Random Field (CRF) from a small number of samples.

In the work by Micilotta et al. [60], the location of a person in the image is found first. Part detectors are learned

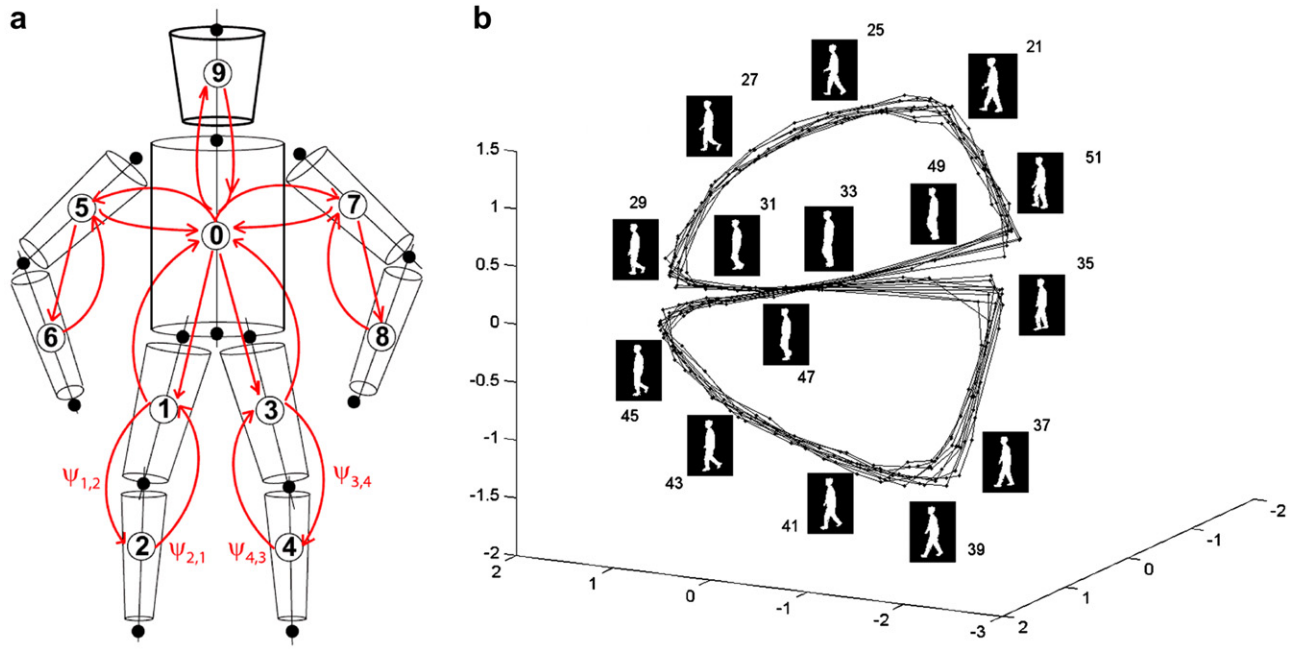

Fig. 3. (a) Relation between body parts described in a graphical model [99] (© MIT Press, 2003); (b) View-based manifold for walking activity [24] (© IEEE, 2004)).

and an assembly is found by applying RANSAC. Heuristics are used to filter unlikely poses, and a pose prior determines the likelihood of the assembly. An example-based approach (see also Section 4.2) is used to find the most likely pose based on extracted silhouette, edges, and hand locations. Although this approach is computationally very efficient, only frontal poses are regarded. It would be interesting to see how the work could be generalized to more unconstrained movements.

3.1.3. Combined top–down and bottom–up estimation

By combining pure top–down and bottom–up approaches, the drawbacks of both can be targeted. Automatic initialization can be achieved within a sound tracking framework.

Navaratnam et al. [70] use a search-space decomposition approach. Body parts lower in the kinematic chain are found using part detectors within an image region that is defined by the parent in the kinematic chain. This approach is computationally less expensive but performance depends heavily on the individual part detectors. Demirdjian [20] uses optical flow in a top–down approach to select a candidate pose estimate. In addition, a view-based key frame that describes the appearance of the person is selected. The motion between the support points of the key frame and the image is used to refine the estimate. The final pose estimate is obtained by fusing both model-based and view-based estimates.

Hua et al. [39] incorporate bottom–up information in a statistical framework. Comparable to Sigal et al. [99], the human body is modeled as a Markov network. 2D body poses are inferred using a data driven belief propagation Monte Carlo algorithm. Shape, edge and color cues are used to construct the importance sampling functions.

Lee et al. [54] use part detectors and inverse kinematics to estimate part of the pose space. Bottom–up information is only used when available, eliminating the need for a part detector for each limb. The approach targets the drawbacks of a pure top–down approach, while still providing a flexible tracking framework. However, the bottom–up information in used in a fixed analytical way. Not only does this approach require fixed segment lengths, it also prevents correct estimation of certain types of poses (e.g., poses where the elbow is higher than the hand). In [53], proposal maps are introduced to facilitate the mapping from 2D observations to 3D pose space.

Recent work has focussed on the recovery of human poses in cluttered scenes. [55] adopt a three-stage approach, based on [53], to subsequently find human bodies, their 2D body part locations and a 3D pose estimate. Sminchisescu et al. [103] learn top–down and bottom–up functions in alternate steps. The bottom–up process is tuned using samples from the top–down process, which is optimized to produce estimates that are close to those predicted by the bottom–up process. The processes are guaranteed to converge to equilibrium.

3.2. Single and multiple hypothesis tracking

Estimating poses from frame to frame is usually termed tracking. Tracking is used to ensure temporal coherence between poses over time, and to provide an initial pose estimate. When it is assumed that the time between subsequent frames is small, the distance in body configuration is likely to be small as well. These configuration differences can be approximately linearly tracked, for example using a Kalman filter. Traditional tracking was aimed at maintaining

a single hypothesis over time. Since this often causes the estimation to lose track, most recent work propagates multiple hypothesis in time. Often, a sampling-based approach is taken. In some works, temporal coherence is achieved by minimizing pose changes over a sequence of frames in a batch approach. Related to this is the estimation of 3D poses from 2D points. Although this topic is outside the scope of our overview, it is relevant and we choose to include it. This section discusses these methodologies.

3.2.1. Single hypothesis tracking

The high dimensionality of the pose space prohibits an exhaustive search of the cost surface. Single hypothesis approaches include Kalman filtering and local-optimization methods [13,118,45]. Gavrila and Davis [28] use a discrete estimation to reduce computation time.

Single hypothesis tracking suffer from accumulation of errors. In case of ambiguity, such as self-occlusion, there is always the possibility of selecting the wrong pose. By maintaining only a single hypothesis, the pose estimation is likely to 'drift off' which makes recovery difficult.

3.2.2. Multiple hypothesis tracking

To overcome the drifting problem of single hypothesis tracking approaches, multiple hypotheses can be maintained. Cham and Rehg [16] use a set of Kalman filters to propagate multiple hypotheses. This results in more reliable motion tracking than with a single Kalman filter. Evaluation on challenging dancing sequences shows that the multiple hypotheses are able to track movement where a single mode fails. However, due to their limited appearance model, rotations about limb axes could not be estimated.

Human motion is non-linear due to joint accelerations. However, Kalman filters are only suitable for tracking linear motion. Sampling-based approaches (particle filtering or CONDENSATION [29,42]) are able to track non-linear motion. In general, a number of particles is propagated in time using a model of dynamics, including a noise component. Each particle has an associated weight, that is updated according to the cost function. Configurations with a low cost are assigned a high weight. Since all weights sum up to one, the pose estimate is obtained by the weighted sum of all particles. (Or alternatively, the particle with the maximum weight is selected.)

Although, in theory, sampling-based methods are very suitable for tracking, the high dimensionality requires the use of many particles to sample the pose space sufficiently densely. Every particle comes with an increase in computational cost due to propagating the particles according to the dynamical model and the evaluation of the cost function. For each particle, the human body model must be rendered and compared to the extracted image descriptors. Another problem is the fact that particles tend to cluster themselves on a very small area. This is called sample impoverishment [48], and leads to a decreasing number of effective particles. Different particle sampling schemes have been proposed to overcome this problem. In [122], some common schemes are evaluated quantitatively on the human motion tracking task.

Currently, there are two main solutions to make the problem more tractable. The first one is to use priors on the movement that can be recognized. This includes learning motion models to guide the particles more effectively, and to learn a low-dimensional space which reduces the number of particles needed. We discuss these topics in Section 3.4. A second solution is to spread particles more efficiently in places where a suitable local minimum is more likely. We discuss this solution below.

Sminchisescu and Triggs [105] introduce Covariance Scaled Sampling (CSS) to guide the particles. Instead of inflating the noise component in the model of dynamics, the posterior covariance of the previous frame is inflated. Intuitively, this focuses the particles in the regions where there is uncertainty, for example due to depth ambiguities as observed in monocular tracking. In the unconstrained case and given monocular data and known segment length, each joint has a twofold ambiguity. The connected limb is either placed forwards, or backwards. This also means that there are two local minima. When tracking fails, this is most likely due to choosing the wrong minimum. In [106], these ambiguities are enumerated in a tree, and the particles are allowed to 'jump' in the pose space accordingly. Deutscher and Reid [21] introduce a different approach to guide the particles. They use simulated annealing to focus the particles on the global maxima of the posterior, at the price of multiple iterations per frame. Particles are distributed widely at initialization, and their range of movement is decreased gradually over time.

MacCormick and Isard [59] partition the pose space into a number of lower-dimensional subspaces. Because independence between the spaces is assumed, this idea is similar to search-space decomposition. As we discussed in the previous section, Lee et al. [54] avoid the need of an inhibitingly large number of particles by updating part of the state space using analytical inference.

3.2.3. Batch methods

Batch methods optimize poses over a sequence of frames, and are therefore unsuitable for online tracking. They avoid the need of propagating multiple hypotheses, since the most likely sequence of poses can be determined automatically. Plänkers and Fua [79] and Liebowitz and Carlsson [57] use least-squares minimization, Brand [11] and Navaratnam et al. [70] use the Viterbi algorithm to find the most probable state sequence in an Hidden Markov Model (HMM).

3.3. 3D pose estimation from 2D points

When only 2D points over a sequence of images are known, 3D poses can be estimated if a human body model is taken into account. Liebowitz and Carlsson [57] reconstruct 3D poses from 2D point correspondences from multiple views and known body segment lengths. Linear

geometric reconstruction is used to recover the poses of an entire motion sequence at once. Taylor [111] uses only a single view and recovers the entire set of pose solutions by considering the foreshortening of the segments of the model in the image. A scaled orthographic projection is assumed, which limits the approach to far views. Depth ordering must be specified manually. Lee and Chen [52] recover the camera parameters from 6 points on the head. They use an interpretation tree to store all kinematic ambiguities that arise from forward to backward flipping and apply a number of constraints to prune impossible configurations. Additionally, DiFranco et al. [22] use user-specified 3D key frames. A maximum *a posteriori* trajectory is calculated using a non-linear least squares framework, taking into account joint angle limits and smooth dynamics. In [76], no camera model is assumed but fixed segment ratios are used.

3.4. Motion priors

Although the human body can perform a very broad variety of movements, the set of typically performed movements is usually much smaller. Especially when only a single class of movements (e.g., walking, swimming) is regarded, motion priors can aid in performing more stable tracking. However, this comes as a cost of putting a strong restriction on the poses that can be recovered.

Many prior models are derived from training data. A possible weakness of these motion models is that the ability to accurately represent the space of realizable human movements generally depends significantly on the amount of available training data. Therefore, the set of exemplars must be sufficiently large and account for the variations that can be observed while tracking the movement.

Generally, we can identify two main classes of motion priors. The first uses an explicit motion model to guide the tracking. The second class learns a low-dimensional activity manifold, in which tracking occurs.

3.4.1. Using motion models

Most statistical motion models can only be used for specific movements, such as walking [34,87] dancing [83] or tennis [108]. However, more general models exist [1,77,94].

Howe et al. [38] use snippets of motion from a database to recover 3D motion given 2D points. From a sequence of 2D poses, the 3D motion is reconstructed by finding the MAP estimate of the sequence of snippets. Sidenbladh et al. [94] take a similar approach. They retrieve motion samples similar to the motion being tracked. The dynamics of the sample are used to propagate the particles in a particle filter framework. Ning et al. [71] use a similar approach, but constrain the propagation of the particles using physical motion constraints.

Instead of using samples, Pavlovic et al. [77] learn a dynamical model over the pose space. Agarwal and Triggs [1] cluster their training data into body poses with similar dynamics. Principal Component Analysis (PCA) is applied to reduce the dimensionality for each cluster, followed by learning a local linear autoregression. A class inference algorithm is able to estimate the current motion cluster and allows for smooth transitions between classes.

The work of [14] does not only model the short-term dynamics but also takes into account the history using Variable Length Markov Models (VLMM). Clusters of elementary motion are learned from training data and clustered. State transitions in the VLMM correspond to one of the clusters. Particles are propagated according to the dynamics of the selected cluster. The noise vector, added in the propagation, is sampled from the covariance of the cluster. This is similar in spirit to CSS [105], where the noise is sampled from the covariance of the previous posterior distribution.

3.4.2. Dimensionality reduction

Reducing the dimensionality of the pose space is motivated by the observation that human activities are often located on a latent space that is low-dimensional [24,31]. As mentioned before, tracking in this low-dimensional manifold results in lower numbers of required particles. Currently, manifolds are learned for specific activities, such as walking, and it remains to be researched how this can be extended to broader classes of movement.

Tracking in a low-dimensional manifold requires three components. First, a mapping between original pose space to low-dimensional manifold must be learned. Second, an inverse mapping must be defined. Third, it must be defined how tracking within the low-dimensional space occurs.

Since the mapping between the original pose space and latent space is in general non-linear, linear PCA is inadequate. Algorithms such as Locally Linear Embedding and Isomap can learn this non-linear mapping but are not invertible. This inverse mapping is needed because the full body configuration is required for evaluation of the likelihood function. Gaussian Process Latent Variable Models (GPLVM, [51]) and Locally Linear Coordination (LLC, [112]) do provide the inverse mapping.

Sminchisescu and Jepson [101] use spectral embedding to learn the embedding, which is modeled as a Gaussian mixture model. Radial Basis Functions (RBF) are learned for the inverse mapping. A linear dynamical model is used for tracking. Urtasun et al. [116] use a GPLVM to learn prior models for 3D human tracking. GPLVMs generate smooth mappings between pose space and latent space, which is useful for the use of gradient descent to optimize pose estimates. A second-order Gauss–Markov model is used as a motion model. In later work [119,115], a Gaussian Process Dynamical Model (GPDM) is learned from training data. The GPDM also learns a dynamical model in the latent space. Recent work by Moon and Pavlović [65] has investigated the effect of dynamics in the embedding on human motion tracking.

Tian et al. [113] use a GPLVM for 2D pose estimation. Particle filtering is used, where the samples are drawn from the latent space. Alternatively, Li et al. [56] use LLC for learning the mappings. Smoothing in the latent space is

not enforced but the mapping is such that close points in latent space correspond to close poses in the pose space. Therefore, a simple dynamical model can be used.

4. Model-free approaches

If no explicit human body model is available, a direct relation between image observation and pose must be established. Two main classes of pose estimation approach can be identified: learning-based (Section 4.1) and example-based (Section 4.2). In learning-based approaches, a function from image space to pose space is learned using training data. Example-based approaches avoid learning this mapping. Instead, a collection of exemplars is stored in a database, together with their corresponding pose descriptions. For a given input image, a similarity search is performed and candidate poses are interpolated to obtain the pose estimate. Note that although the inverse mapping from image space to pose estimate is multi-valued and cannot be functionally approximated [102], most work treats the relation as single-valued.

Since variations in body configuration, body dimensions, viewpoint and appearance are implicitly modeled in the training data, this data needs to generalize well over the invariant parameters and distinguish well between the variant ones. The training data must account for the high non-linearity of the mapping between image and pose space, which means in practice that the pose space must be densely sampled in the training set. However, the training data can be constructed when keeping in mind that not all kinematically possible poses are also likely.

Model-free algorithms do not suffer from (re)initialization problems and can in this respect be used for initialization of model-based pose estimation approaches as we discussed in Section 3.

4.1. Learning-based

Grauman et al. [30] describe a distribution over both multi-view silhouettes and 3D joint locations with a mixture of probabilistic PCA. Pose inference is based on the maximum *a posteriori* (MAP) estimate. Silhouettes from a single view are used by Agarwal and Triggs [3]. They use non-linear regression to model the relation between histograms of shape contexts and 3D poses. Damped least-squares and Relevance Vector Machine regression over both linear and kernel bases have been evaluated. Ambiguities are resolved using dynamics.

In recent work, Agarwal and Triggs [2] use histograms of gradient orientations over a grid of small cells. Non-negative matrix factorization is used to obtain a set of basis vectors that correspond to local features on the human body such as shoulders and bent elbows. When using these vectors to reconstruct an image with clutter, the edges that correspond to the person are obtained. This enables them to recover poses without having to extract the person's outline. Regression is used to recover upper-body poses.

Brand [11] models a manifold of pose and velocity configurations with an HMM. Temporal ambiguities are resolved by recovering poses over an entire sequence by applying the Viterbi algorithm. Elgammal and Lee [24] recover 3D poses from monocular silhouettes using an intermediary activity manifold (see Fig. 3(b)). Manifolds are learned from visual input and subsequently, mappings are learned from manifolds to visual input and 3D poses. Good generalization for variations in body shape are reported. However, the manifolds are learned for specific activities and viewpoints, and it is unclear how the work would generalize to a more unconstrained motion domain. In [109], a pose manifold is learned in addition to the image manifold. LLE is used to learn a mapping between the two manifolds.

Rosales and Sclaroff [89] observe that the inverse of the mapping from image space to pose space cannot be modeled by a single function. Therefore, they cluster the 2D pose space and learn specialized functions for each cluster from image descriptors to pose space. A neural network is used as mapping function. In [90], the work is extended to allow input from multiple cameras. The pose is estimated for each camera individually and in a subsequent step, the hypotheses are combined into a set of self-consistent 3D pose hypotheses. Sminchisescu et al. [102] model the multi-valued nature of the mapping from observation to pose state with a mixture of expert models. Each expert learns the conditional state distributions from a database consisting of samples of pose representation and a rendered human body model. Shape contexts in addition to local appearance are used as image descriptors. The samples involve a number of human activities such as walking, running and pantomime. Demonstration on monocular complex motions shows convincing results, and tests on artificial data show that the proposed approach outperforms nearest-neighbor and regression methods. Training these mappings requires large amounts of labelled example pairs consisting of both image descriptors and poses. In [69], also data from each of the types separately are used to improve manifold learning.

Recent work by Taycher et al. [110] transforms the continuous state estimation problem into a discrete one by using dividing the state space into regions that approximate the posterior. The observation potential function of the CRF is learned off-line from a large number of examples. By focusing only on the regions where the prior state probability is significant, poses can be recovered in real time.

4.2. Example-based

Example-based approaches use a database of exemplars that describe poses in both image space and pose space. One drawback of these approaches is the large amount of space needed to store the database.

Mori and Malik [66] extract external and internal contours of an object. Shape contexts are employed to encode

the edges. In an estimation step, the stored exemplars are deformed to match the image observation. In this deformation, the location of the hand-labelled 2D locations of joints also changes. The most likely 2D joint estimate is found by enforcing 2D image distance consistency between body parts. Shape deformation is also used by Sullivan and Carlsson [108]. To improve the robustness of the point transferral, the spatial relationship of the body points and color information is exploited. Loy et al. [58] perform interpolation between key frame poses based on [111] and additional smoothing constraints. Manual intervention is necessary in certain cases.

Bowden et al. [10] fit a non-linear point distribution model (PDM) to their image observations. The PDM consists of the 2D position of head and hands in the image, the 2D body contour, and the 3D structure of the body. The PDM is trained on high-dimensional feature vectors that contain likely body movements. The feature space is projected on a lower dimensional space. In [75], the poses in the database are rendered from multiple views, which makes the approach somewhat invariant to the viewpoint. For a monocular image, the view is estimated using a linear discriminant and subsequently the pose is recovered using a nearest neighbor classifier. Ong and Gong [72] include views from multiple cameras in the PDM and recover a pose from multi-view images.

Toyama and Blake [114] also show how to incorporate exemplars in a probabilistic temporal framework. Silhouettes, described using turning angle and Chamfer distance are considered by Howe [35]. To achieve temporal coherence, he uses Markov Chaining with subsequent smoothing over a sequence of frames. In later work [36], optical flow information is used in addition. Motion is used in the estimation process by Ong et al. [73]. Their exemplar space is clustered and flow vectors between clusters are learned from sequences of training data. A particle filter framework is used where the particles are guided by the flow vectors. This reduces the number of particles needed but puts a strong prior on the motions that can be estimated.

The computational complexity of a naive Nearest Neighbor search is linear in the number of exemplars. For recovering more unconstrained movements or high number of DOF, the number of exemplars grows substantially. Therefore, Shakhnarovich et al. [91] introduce Parameter Sensitive Hashing (PSH) to rapidly estimate the pose given a new image. Because of the ambiguity in the use of silhouettes alone, they use edge direction histograms within a contour. PSH is also applied in [83], where a bit string of binary local features [117] extracted from silhouettes obtained using three views are used instead. In addition to PSH, they use a motion graph to find those poses that are not only close in image space, but are also close in pose space.

5. Discussion

Human motion analysis is a challenging problem due to large variations in human motion and appearance, camera viewpoint and environment settings. On the other hand, we know much about people's physical appearance and movements. The key point for successful human motion analysis is to use this knowledge effectively. Over the last two decades, a large amount of research has been conducted. Human body models that were initially described in 2D have now evolved into highly articulated 3D models. Deterministic linear tracking has been replaced by sampling-based tracking frameworks that evaluate the cost function effectively. The role of machine learning plays an increasingly important role in human motion analysis, and will continue to do so.

For each of the methodologies described in this survey, prior knowledge about human movement or appearance is incorporated more and more effectively. For example, joint angle limitations are directly encoded during tracking, instead of as a pose space pruning technique. But although many of these advances have led to impressive results given the complexity of the task, the domain was always limited. Not unusually, it is assumed that a person has been found in the image in a preprocessing step. Furthermore, assumptions about the viewpoint, appearance and motion are often made.

We expect that combining methodologies is the solution to use prior knowledge even more effectively. Indeed, recent work explores these kind of combinations. While much research is needed, these works are certainly promising. For example, model-based and model-free approaches have been combined [60] to allow for automatic initialization and recovery. Another promising direction of research is the recent combination of bottom–up and top–down approaches, as described in Section 3.1. This has led to effective tracking frameworks. Also, 2D and 3D models have been combined to facilitate detection and subsequent pose estimation [3,12]. Also, they have the potential to deal more effectively with occlusions, a problem that is often ignored. Work by Howe [37] also addresses this issue.

Also, the role of context should be used more explicitly. Human motion analysis provides input for reasoning about actions and intentions. Reversely, context can be used for human motion analysis, other than implicitly by assuming a fixed domain. Recent work aims at learning models that are conditioned on the context [14,104].

The role of human motion models, and how they generalize to broader domains remains to be investigated. Also, the suitability of low-dimensional latent spaces for recovery of more spontaneous movement needs to be assessed.

From a practical perspective, evaluation of motion analysis algorithms requires a common database, representative for a broad range of domains (indoor, static scenes, and dynamic, cluttered scenes with multiple persons). This database should consist of ground truth data and image sequences. In addition, common criteria (accuracy, smoothness, speed) for evaluation are needed. The recently introduced HumanEva-I database [96] is a good first step in this direction. When the evaluation criteria are generally accepted, this will contribute significantly in determining promising directions of research.

Acknowledgments

This work was supported by the European IST Programme Project FP6-033812 (Augmented Multi-party Interaction with Distant Access, publication AMIDA-3), and is part of the ICIS program. ICIS is sponsored by the Dutch government under contract BSIK03024. The author wishes to thank Dariu Gavrila and the anonymous CVIU reviewers for their valuable comments, and all authors that contributed figures to this overview.

References

[1] Ankur Agarwal, Bill Triggs, Tracking articulated motion using a mixture of autoregressive models, in: Proceedings of the European Conference on Computer Vision (ECCV'04), Lecture Notes in Computer Science, vol. 3 (3024), Prague, Czech Republic, May 2004, pp. 54–65.

[2] Ankur Agarwal, Bill Triggs, A local basis representation for estimating human pose from cluttered images, in: Proceedings of the Asian Conference on Computer Vision (ACCV'06)—Part 1, Lecture Notes in Computer Science, vol. 3851, Hyderabad, India, January 2006, pp. 50–59.

[3] Ankur Agarwal, Bill Triggs, Recovering 3D human pose from monocular images, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 28 (1) (2006) 44–58.

[4] Jake K. Aggarwal, Qin Cai, Human motion analysis: a review, Computer Vision and Image Understanding (CVIU) 73 (3) (1999) 428–440.

[5] Carlos Barrón, Ioannis A. Kakadiaris, Estimating anthropometry and pose from a single uncalibrated image, Computer Vision and Image Understanding (CVIU) 81 (3) (2001) 269–284.

[6] Carlos Barrón, Ioannis A. Kakadiaris, Monocular human motion tracking, Multimedia Systems 10 (2004) 118–130.

[7] Serge Belongie, Jitendra Malik, Jan Puzicha, Shape matching and object recognition using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 24 (4) (2002) 509–522.

[8] Chiraz BenAbdelkader, Larry S. Davis, Estimation of anthropomeasures from a single calibrated camera, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR'06), Southampton, United Kingdom, April 2006, pp. 499–504.

[9] Andrea Bottino, Aldo Laurentini, A silhouette-based technique for the reconstruction of human movement, Computer Vision and Image Understanding (CVIU) 83 (1) (2001) 79–95.

[10] Richard Bowden, Tom A. Mitchell, Mansoor Sarhadi, Non-linear statistical models for the 3D reconstruction of human pose and motion from monocular image sequences, Image and Vision Computing 18 (9) (2000) 729–737.

[11] Matthew Brand, Shadow puppetry, in: Proceedings of the International Conference on Computer Vision (ICCV'99), vol. 2, Kerkyra, Greece, September 1999, pp. 1237–1244.

[12] Matthieu Bray, Pushmeet Kohli, Philip H. Torr, Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts, in: Proceedings of the European Conference on Computer Vision (ECCV'06), Lecture Notes in Computer Science, vol. 2 (3952), Graz, Austria, May 2006, pp. 642–655.

[13] Christoph Bregler, Jitendra Malik, Katherine Pullen, Twist based acquisition and tracking of animal and human kinematics, International Journal of Computer Vision 56 (3) (2004) 179–194.

[14] Fabrice Caillette, Aphrodite Galata, Toby Howard, Real-time 3-D human body tracking using variable length markov models, in: Proceedings of the British Machine Vision Conference (BMVC'05), vol. 1, Oxford, United Kingdom, September 2005, pp. 469–478.

[15] Joel Carranza, Christian Theobalt, Marcus A. Magnor, Hans-Peter Seidel, Free-viewpoint video of human actors, ACM Transactions on Computer Graphics 22 (3) (2003) 569–577.

[16] Tat-Jen Cham, James M. Rehg, A multiple hypothesis approach to figure tracking, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'99), vol. 2, Ft. Collins, CO, June 1999, pp. 239–245.

[17] German K.M. Cheung, Simon Baker, Takeo Kanade, Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'03), vol. 1, Madison, WI, June 2003, pp. 77–84.

[18] Chi-Wei Chu, Odest C. Jenkins, Maja J. Mataric, Markerless kinematic model and motion capture from volume sequences, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'03), vol. 2, Madison, WI, June 2003, pp. 475–483.

[19] Quentin Delamarre, Olivier Faugeras, 3D articulated models and multiview tracking with physical forces, Computer Vision and Image Understanding (CVIU) 81 (3) (2001) 328–357.

[20] David Demirdjian, Combining geometric- and view-based approaches for articulated pose estimation, in: Proceedings of the European Conference on Computer Vision (ECCV'04), Lecture Notes in Computer Science, vol. 3 (3023), Prague, Czech Republic, May 2004, pp. 183–194.

[21] Jonathan Deutscher, Ian Reid, Articulated body motion capture by stochastic search, International Journal of Computer Vision 61 (2) (2005) 185–205.

[22] David E. DiFranco, Tat-Jen Cham, James M. Rehg, Reconstruction of 3-D figure motion from 2-D correspondences, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'01), vol. 1, Kauai, HI, December 2001, pp. 307–314.

[23] Tom Drummond, Roberto Cipolla, Real-time tracking of highly articulated structures in the presence of noisy measurements, in: Proceedings of the International Conference On Computer Vision (ICCV'01), vol. 2, Vancouver, Canada, July 2001, pp. 315–320.

[24] Ahmed M. Elgammal, Chan-Su Lee, Inferring 3D body pose from silhouettes using activity manifold learning, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'04), vol. 2, Washington, DC, June 2004, pp. 681–688.

[25] Ali Erol, George Bebis, Mircea Nicolescu, Richard D. Boyle, Xander Twombly, Vision-based hand pose estimation: A review, Computer Vision and Image Understanding, this issue, doi:10.1016/j.cviu.2006.10.012.

[26] Pedro F. Felzenszwalb, Daniel P. Huttenlocher, Pictorial structures for object recognition, International Journal of Computer Vision 61 (1) (2005) 55–79.

[27] Dariu M. Gavrila, The visual analysis of human movement: A survey, Computer Vision and Image Understanding (CVIU) 73 (1) (1999) 82–92.

[28] Dariu M. Gavrila, Larry S. Davis, Tracking of humans in action: A 3D model-based approach, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'96), San Francisco, CA, June 1996, pp. 73–80.

[29] Neil J. Gordon, David J. Salmond, Adrian F.M. Smith, Novel approach to nonlinear/nonGaussian Bayesian state estimation, in: IEE Proceedings-F (Radar and Signal Processing), vol. 140, April 1993, pp. 107–113.

[30] Kristen Grauman, Gregory Shakhnarovich, Trevor Darrell, Inferring 3D structure with a statistical image-based shape model, in: Proceedings of the International Conference on Computer Vision (ICCV'03), vol. 1, Nice, France, October 2003, pp. 641–647.

[31] Keith Grochow, Steven L. Martin, Aaron Hertzmann, Zoran Popovic, Style-based inverse kinematics, ACM Transactions on Graphics 23 (3) (2004) 522–531.

[32] Tony X. Han, Huazhong Ning, Thomas S. Huang, Efficient nonparametric belief propagation with application to articulated body tracking, in: Proceedings of the Conference on Computer

Vision and Pattern Recognition (CVPR'06), vol. 1, New York, NY, June 2006, pp. 214–221.
[33] Ismail Haritaoglu, David Harwood, Larry S. Davis, W^4s: A real-time system detecting and tracking people in 2 1/2D, in: Proceedings of the European Conference on Computer Vision (ECCV'98), Lecture Notes in Computer Science, vol. 1 (1406), Freiburg, Germany, June 1998, pp. 877–892.
[34] David Hogg, Model-based vision: a program to see a walking person, Image and Vision Computing 1 (1) (1983) 5–20.
[35] Nicholas R. Howe, Silhouette lookup for automatic pose tracking, in: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04), Los Alamitos, CA, June 2004, p. 15.
[36] Nicholas R. Howe, Flow lookup and biological motion perception, in: Proceedings of the Internation Conference on Image Processing (ICIP'05), vol. 3, Genova, Italy, September 2005, pp. 1168–1171.
[37] Nicholas R. Howe, Boundary fragment matching and articulated pose under occlusion, in: Proceedings of the International Conference on Articulated Motion and Deformable Objects (AMDO'06), Lecture Notes in Computer Science, (4069), Port d'Andratx, Spain, July 2006, pp. 271–280.
[38] Nicholas R. Howe, Michael E. Leventon, William T. Freeman, Bayesian reconstruction of 3D human motion from single-camera video, in: Advances in Neural Information Processing Systems (NIPS) 12, Denver, CO, November 2000, pp. 820–826.
[39] Gang Hua, Ming-Hsuan Yang, Ying Wu, Learning to estimate human pose with data driven belief propagation, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 2, San Diego, CA, June 2005, pp. 747–754.
[40] Yu Huang, Thomas S. Huang, Model-based human body tracking, in: Proceedings of the International Conference on Pattern Recognition (ICPR'02), vol. 1, Quebec, Canada, August 2002, pp. 552–555.
[41] Sergey Ioffe, David A. Forsyth, Probabilistic methods for finding people, International Journal of Computer Vision 43 (1) (2001) 45–68.
[42] Michael Isard, Andrew Blake, CONDENSATION—conditional density propagation for visual tracking, International Journal of Computer Vision 29 (1) (1998) 5–28.
[43] Nebojsa Jojic, Jin Gu, Helen Shen, Thomas S. Huang, 3-D reconstruction of multipart, self-occluding objects, in: Proceedings of the Asian Conference on Computer Vision (ACCV'98), Hong Kong, China, January 1998, pp. 455–462.
[44] Shanon X. Ju, Michael J. Black, Yaser Yacoob, Cardboard people: A parameterized model of articulated image motion, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR'96), Killington, VT, October 1996, pp. 38–44.
[45] Ioannis A. Kakadiaris, Dimitris N. Metaxas, Three-dimensional human body model acquisition from multiple views, International Journal of Computer Vision 30 (3) (1998) 191–218.
[46] Ioannis A. Kakadiaris, Dimitris N. Metaxas, Model-based estimation of 3D human motion, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 22 (12) (2000) 1453–1459.
[47] Roland Kehl, Luc Van Gool, Markerless tracking of complex human motions from multiple views, Computer Vision and Image Understanding (CVIU) 104 (2–3) (2006) 190–209.
[48] Oliver D. King, David A. Forsyth, How does CONDENSATION behave with a finite number of samples?, in: Proceedings of the European Conference on Computer Vision (ECCV'00), Lecture Notes in Computer Science, vol. 1 (1842), Dublin, Ireland, June 2000, pp. 695–709.
[49] Nils Krahnstöver, Mohammed Yeasin, Rajeev Sharma, Automatic acquisition and initialization of articulated models, Machine Vision and Applications 14 (4) (2003) 218–228.
[50] Xiangyang Lan, Daniel P. Huttenlocher, Beyond trees: common-factor models for 2D human pose recovery, in: Proceedings of the International Conference On Computer Vision (ICCV'05), vol. 1, Beijing, China, October 2005, pp. 470–477.
[51] Neil D. Lawrence, Gaussian process latent variable models for visualisation of high dimensional dataAdvances in Neural Information Processing Systems (NIPS), vol. 16, Vancouver, Canada, 2003, pp. 329–336.
[52] Hsi-Jian J. Lee, Zen Chen, Determination of 3D human body posture from a single view, Computer Vision, Graphics and Image Processing 30 (2) (1985) 148–168.
[53] Mun Wai Lee, Isaac Cohen, Proposal maps driven mcmc for estimating human body pose in static images, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'04), vol. 2, Washington, DC, June 2004, pp. 334–341.
[54] Mun Wai Lee, Isaac Cohen, Soon Ki Jung, Particle filter with analytical inference for human body tracking, in: Proceedings of the Workshop on Motion and Video Computing (MOTION'02), Orlando, FL, December 2002, pp. 159–168.
[55] Mun Wai Lee, Ramakant Nevatia, Human pose tracking using multi-level structured models, in: Proceedings of the European Conference on Computer Vision (ECCV'06), Lecture Notes in Computer Science, vol. 3 (3953), Graz, Austria, May 2006, pp. 368–381.
[56] Rui Li, Ming-Hsuan Yang, Stan Sclaroff, Tai-Peng Tian, Monocular tracking of 3D human motion with a coordinated mixture of factor analyzers, in: Proceedings of the European Conference on Computer Vision (ECCV'06), Lecture Notes in Computer Science, vol. 2 (3952), Graz, Austria, May 2006, pp. 137–150.
[57] David Liebowitz, Stefan Carlsson, Uncalibrated motion capture exploiting articulated structure constraints, International Journal of Computer Vision 51 (3) (2003) 171–187.
[58] Gareth Loy, Martin Eriksson, Josephine Sullivan, Stefan Carlsson, Monocular 3D reconstruction of human motion in long action sequences, in: Proceedings of the European Conference on Computer Vision (ECCV'04), Lecture Notes in Computer Science, vol. 4 (3024), Prague, Czech Republic, May 2004, pp. 442–455.
[59] John MacCormick, Michael Isard, Partitioned sampling, articulated objects, and interface-quality hand tracking, in: Proceedings of the European Conference on Computer Vision (ECCV'00), Lecture Notes in Computer Science, vol. 2 (1843), Dublin, Ireland, June 2000, pp. 3–19.
[60] Antonio S. Micilotta, Eng-Jon Ong, Richard Bowden, Real-time upper body detection and 3D pose estimation in monoscopic images, in: Proceedings of the European Conference on Computer Vision (ECCV'06), Lecture Notes in Computer Science, vol. 3 (3953), Graz, Austria, May 2006, pp. 139–150.
[61] Ivana Mikić, Mohan Trivedi, Edward Hunter, Pamela Cosman, Human body model acquisition and tracking using voxel data, International Journal of Computer Vision 53 (3) (2003) 199–223.
[62] Anurag Mittal, Liang Zhao, Larry S. Davis, Human body pose estimation using silhouette shape analysis, in: Proceedings of the Conference on Advanced Video and Signal Based Surveillance (AVSS'03), Miami, FL, July 2003, pp. 263–270.
[63] Thomas B. Moeslund, Erik Granum, A survey of computer vision-based human motion capture, Computer Vision and Image Understanding (CVIU) 81 (3) (2001) 231–268.
[64] Thomas B. Moeslund, Adrian Hilton, Volker Krüger, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding (CVIU) 104 (2–3) (2006) 90–126.
[65] Kooksang Moon, Vladimir I. Pavlović, Impact of dynamics on subspace embedding and tracking of sequences, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1, New York, NY, June 2006, pp. 198–205.
[66] Greg Mori, Jitendra Malik, Recovering 3D human body configurations using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 28 (7) (2006) 1052–1062.
[67] Greg Mori, Xiaofeng Ren, Alexei A. Efros, Jitendra Malik, Recovering human body configurations: Combining segmentation and recognition, in: Proceedings of the Conference on Computer

Vision and Pattern Recognition (CVPR'04), vol. 2, Washington, DC, June 2004, pp. 326–333.
[68] Daniel D. Morris, James M. Rehg, Singularity analysis for articulated object tracking, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'98), Santa Barbara, CA, June 1998, pp. 289–297.
[69] Ramanan Navaratnam, Andrew W. Fitzgibbon, Roberto Cipolla, Semi-supervised learning of joint density models for human pose estimation, in: Proceedings of the British Machine Vision Conference (BMVC'06), vol. 2, Edinburgh, United Kingdom, September 2006, pp. 679–688.
[70] Ramanan Navaratnam, Arasanathan Thayananthan, Philip H. Torr, Roberto Cipolla, Hierarchical part-based human body pose estimation, in: Proceedings of the British Machine Vision Conference (BMVC'05), Oxford, United Kingdom, September 2005.
[71] Huazhong Ning, Tieniu Tan, Liang Wang, Weiming Hu, People tracking based on motion model and motion constraints with automatic initialization, Pattern Recognition 37 (7) (2004) 1423–1440.
[72] Eng-Jon Ong, Shaogang Gong, A dynamic 3D human model using hybrid 2D-3D representations in hierarchical pca space, in: Proceedings of the British Machine Vision Conference (BMVC'99), Nottingham, United Kingdom, September 1999, pp. 33–42.
[73] Eng-Jon Ong, Antonio S. Micilotta, Richard Bowden, Adrian Hilton, Viewpoint invariant exemplar-based 3D human tracking, Computer Vision and Image Understanding (CVIU) 104 (2–3) (2006) 178–189.
[74] Joseph O'Rourke, Norman I. Badler, Model-based image analysis of human motion using constraint propagation, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 2 (6) (1980) 522–536.
[75] Carlos Orrite-Uruñuela, Jesús Martínez del Rincón, José Elías Herrero-Jaraba, Grégory Rogez, 2D silhouette and 3D skeletal models for human detection and tracking, in: Proceedings of the International Conference on Pattern Recognition (ICPR'04), vol. 4, Cambridge, United Kingdom, August 2004, pp. 244–247.
[76] Vasu Parameswaran, Rama Chellappa, View independent human body pose estimation from a single perspective image, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'04), vol. 2, Washington, DC, June 2004, pp. 16–22.
[77] Vladimir I. Pavlović, James M. Rehg, Tat-Jen Cham, Kevin P. Murphy, A dynamic Bayesian network approach to figure tracking using learned dynamic models, in: Proceedings of the International Conference on Computer Vision (ICCV'99), vol. 1, Kerkyra, Greece, September 1999, pp. 94–101.
[78] Vladimir I. Pavlović, Rajeev Sharma, Thomas S. Huang, Visual interpretation of hand gestures for human–computer interaction: A review, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 19 (7) (1997) 677–695.
[79] Rolf Plänkers, Pascal Fua, Tracking and modeling people in video sequences, Computer Vision and Image Understanding (CVIU) 81 (3) (2001) 285–302.
[80] Deva Ramanan, Learning to parse images of articulated bodies, in: Advances in Neural Information Processing Systems (NIPS) 19, Vancouver, Canada, December 2006, to appear.
[81] Deva Ramanan, David A. Forsyth, Finding and tracking people from the bottom up, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'03), vol. 2, Madison, WI, June 2003, pp. 467–474.
[82] Deva Ramanan, Cristian Sminchisescu, Training deformable models for localization, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1, New York, NY, June 2006, pp. 206–213.
[83] Liu Ren, Gregory Shakhnarovich, Jessica K. Hodgins, Hanspeter Pfister, Paul A. Viola, Learning silhouette features for control of human motion, ACM Transactions on Computer Graphics 24 (4) (2005) 1303–1331.
[84] Xiaofeng Ren, Alexander C. Berg, Jitendra Malik, Recovering human body configurations using pairwise constraints between parts, in: Proceedings of the International Conference On Computer Vision (ICCV'05), vol. 1, Beijing, China, October 2005, pp. 824–831.
[85] Timothy J. Roberts, Stephen J. McKenna, Ian W. Ricketts, Human tracking using 3d surface colour distributions, Image and Vision Computing 24 (12) (2006) 1332–1342.
[86] Grégory Rogez, José J. Guerrero, Jesús Martínez, Carlos Orrite-Uruñuela, Viewpoint independent human motion analysis in man-made environments, in: Proceedings of the British Machine Vision Conference (BMVC'06), vol. 2, Edinburgh, United Kingdom, September 2006, pp. 659–668.
[87] Karl Rohr, Towards model-based recognition of human movements in image sequences, Computer Vision, Graphics, and Image Processing: Image Understanding 59 (1) (1994) 94–115.
[88] Rémi Ronfard, Cordelia Schmid, Bill Triggs, Learning to parse pictures of people, in: Proceedings of the European Conference on Computer Vision (ECCV'02), Lecture Notes in Computer Science, vol. 4 (2353), Copenhagen, Denmark, May 2002, pp. 700–714.
[89] Rómer E. Rosales, Stan Sclaroff, Inferring body pose without tracking body parts, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'00), vol. 2, Hilton Head Island, SC, June 2000, pp. 721–727.
[90] Rómer E. Rosales, Matheen Siddiqui, Jonathan Alon, Stan Sclaroff, Estimating 3D body pose using uncalibrated cameras, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'01), vol. 1, Kauai, HI, December 2001, pp. 821–827.
[91] Gregory Shakhnarovich, Paul A. Viola, Trevor Darrell, Fast pose estimation with parameter-sensitive hashing, in: Proceedings of the International Conference on Computer Vision (ICCV'03), vol. 2, Nice, France, October 2003, pp. 750–759.
[92] Hedvig Sidenbladh, Michael J. Black, Learning the statistics of people in images and video, International Journal of Compututer Vision 54 (1–3) (2003) 181–207.
[93] Hedvig Sidenbladh, Michael J. Black, David J. Fleet, Stochastic tracking of 3D human figures using 2D image motion, in: Proceedings of the European Conference on Computer Vision (ECCV'00), Lecture Notes in Computer Science, vol. 2 (1843), Dublin, Ireland, June 2000, pp. 702–718.
[94] Hedvig Sidenbladh, Michael J. Black, Leonid Sigal, Implicit probabilistic models of human motion for synthesis and tracking, in: Proceedings of the European Conference on Computer Vision (ECCV'02), Lecture Notes in Computer Science, vol. 1 (2350), Copenhagen, Denmark, May 2002, pp. 784–800.
[95] Leonid Sigal, Sidharth Bhatia, Stefan Roth, Michael J. Black, Michael Isard, Tracking loose-limbed people, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'04), vol. 1, Washington, DC, June 2004, pp. 421–428.
[96] Leonid Sigal, Michael J. Black, Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion, Technical Report CS-06-08, Brown University, Department of Computer Science, Providence, RI, September 2006.
[97] Leonid Sigal, Michael J. Black, Measure locally, reason globally: Occlusion-sensitive articulated pose estimation, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, New York, NY, June 2006, pp. 2041–2048.
[98] Leonid Sigal, Michael J. Black, Predicting 3D people from 2D pictures. In Proceedings of the International Conference on Articulated Motion and Deformable Objects (AMDO'06), Lecture Notes in Computer Science, (4069), Port d'Andratx, Spain, July 2006, pp. 185–195.
[99] Leonid Sigal, Michael Isard, Benjamin Sigelman, Michael J. Black, Attractive people: Assembling loose-limbed models using non-parametric belief propagationAdvances in Neural Information Processing Systems (NIPS), vol. 16, Vancouver, Canada, 2003, pp. 1539–1546.
[100] Cristian Sminchisescu, Estimation Algorithms For Ambiguous Visual Models—Three Dimensional Human Modeling And Motion

Reconstruction in: Monocular Video Sequences. PhD thesis, Institute National Politechnique de Grenoble (INPG), Grenoble, July 2002.

[101] Cristian Sminchisescu, Allan D. Jepson, Generative modeling for continuous non-linearly embedded visual inference, in: Proceedings of the International Conference on Machine Learning (ICML'04), Banff, Canada, July 2004, pp. 759–766.

[102] Cristian Sminchisescu, Atul Kanaujia, Zhiguo Li, Dimitris N. Metaxas, Discriminative density propagation for 3D human motion estimation, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, San Diego, CA, June 2005, pp. 390–397.

[103] Cristian Sminchisescu, Atul Kanaujia, Dimitris Metaxas, Learning joint top–down and bottom–up processes for 3D visual inference, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, New York, NY, June 2006, pp. 1743–1752.

[104] Cristian Sminchisescu, Atul Kanaujia, Dimitris N. Metaxas, Conditional models for contextual human motion recognition, Computer Vision and Image Understanding (CVIU) 104 (2–3) (2006) 210–220.

[105] Cristian Sminchisescu, Bill Triggs, Estimating articulated human motion with covariance scaled sampling, International Journal of Robotic Research 22 (6) (2003) 371–392.

[106] Cristian Sminchisescu and Bill Triggs, Kinematic jump processes for monocular 3D human tracking, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'03), vol. 1, Madison, WI, June 2003, pp. 69–76.

[107] Yang Song, Luis Goncalves, Pietro Perona, Unsupervised learning of human motion, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 25 (7) (2003) 814–827.

[108] Josephine Sullivan, Stefan Carlsson, Recognizing and tracking human action, in: Proceedings of the European Conference on Computer Vision (ECCV'02), Lecture Notes in Computer Science, vol. 1 (2350), Copenhagen, Denmark, May 2002, pp. 629–644.

[109] Therdsak Tangkuampien, David Suter, Real-time human pose inference using kernel principal component pre-image approximations, in: Proceedings of the British Machine Vision Conference (BMVC'06), vol. 2, Edinburgh, United Kingdom, September 2006, pp. 599–608.

[110] Leonid Taycher, Gregory Shakhnarovich, David Demirdjian, Trevor Darrell, Conditional random people: Tracking humans with crfs and grid filters, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1, New York, NY, June 2006, pp. 222–229.

[111] Camillo J. Taylor, Reconstruction of articulated objects from point correspondences in a single uncalibrated image, Computer Vision and Image Understanding (CVIU) 80 (3) (2000) 349–363.

[112] Yee Whye Teh, Sam T. Roweis, Automatic alignment of local representationsAdvances in Neural Information Processing Systems (NIPS), vol. 15, Vancouver, Canada, 2002, pp. 841–848.

[113] Tai-Peng Tian, Rui Li, Stan Sclaroff, Tracking human body pose on a learned smooth space. Technical Report BUCS-TR-2005-029, Boston University, Computer Science Department, Boston, MA, July 2005.

[114] Kentaro Toyama, Andrew Blake, Probabilistic tracking with exemplars in a metric space, International Journal of Computer Vision 48 (1) (2002) 9–19.

[115] Raquel Urtasun, David J. Fleet, Pascal Fua, 3D people tracking with gaussian process dynamical models, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1, New York, NY, June 2006, pp. 238–245.

[116] Raquel Urtasun, David J. Fleet, Aaron Hertzmann, Pascal Fua, Priors for people tracking from small training sets, in: Proceedings of the International Conference On Computer Vision (ICCV'05), vol. 1, Beijing, China, October 2005, pp. 403–410.

[117] Paul A. Viola, Michael J. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'01), vol. 1, Kauai, HI, December 2001, pp. 511–518.

[118] Stefan Wachter, Hans-Hellmut Nagel, Tracking persons in monocular image sequences, Computer Vision and Image Understanding (CVIU) 74 (3) (1999) 174–192.

[119] Jack M. Wang, David J. Fleet, Aaron Hertzmann, Gaussian process dynamical modelsAdvances in Neural Information Processing Systems (NIPS), vol. 18, Vancouver, Canada, 2005, pp. 1441–1448.

[120] Jessica J. Wang, Sameer Singh, Video analysis of human dynamics: a survey, Real-Time Imaging 9 (5) (2003) 321–346.

[121] Liang Wang, Weiming Hu, Tieniu Tan, Recent developments in human motion analysis, Pattern Recognition 36 (3) (2003) 585–601.

[122] Ping Wang, James M. Rehg, A modular approach to the analysis and evaluation of particle filters for figure tracking, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1, New York, NY, June 2006, pp. 790–797.

[123] Christopher R. Wren, Ali J. Azarbayejani, Trevor Darrell, Alex P. Pentland, Pfinder: Real-time tracking of the human body, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 19 (7) (1997) 780–785.

[124] Masanobu Yamamoto, Katsutoshi Yagishita, Scene constraints-aided tracking of human body, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'00), vol. 1, Hilton Head Island, SC, June 2000, pp. 151–156.

[125] Wen-Yi Zhao, Rama Chellappa, P. Jonathon Phillips, Azriel Rosenfeld, Face recognition: A literature survey, ACM Computing Surveys 35 (3) (2003) 399–458.