Single Camera 3D Human Pose Estimation: A Review of Current Techniques

Yap Wooi Hen
Dept. of Electrical Engineering
University of Malaya
Lembah Pantai, 50603 Kuala Lumpur.

Raveendran Paramesran
Dept. of Electrical Engineering
University of Malaya
Lembah Pantai, 50603 Kuala Lumpur.

***Abstract*- Vision-based techniques to estimate 3D human pose from single camera are important in many applications like entertainment and games industries, sports science, gait analysis and video surveillance. While de-facto optical-based motion capture system is highly accurate in estimating 3D human pose, vision-based techniques offer better alternative because the latter are far cheaper and non-intrusive. In this paper, we review recent significant advances made in single-camera 3D human pose estimation techniques. We discuss about challenges faced, typical image observations used, ways to address the challenges and highlighting limitations in state-of-the-art. We conclude this paper with speculation of future research directions as well as open research problems.**

I. INTRODUCTION

Techniques to estimate 3D human pose using vision-based system have many important applications. In these applications, it is common that 3D human pose estimation are part of the building blocks, hence its accuracy greatly affects the performance of the applications. For instance, in entertainment and computer games industries, estimated 3D human pose is used to create realistic animation of human movements. In sports science, accurate reconstruction of 3D human pose assists athletes to visually analyse their movements to enhance their movement performance. In physiotherapy field, gait analysis uses 3D human pose to identify the underlying causes of patient's movement anomalies that maybe caused by stroke, cerebral palsy or other neuromuscular problems. Another emerging application of pose estimation is in the area of video surveillance. With lower costs of cameras and advances in computing power, accurate analysis of 3D human pose from video can help surveillance operators to identify events such as running, walking, shop-lifting, wall-climbing, loitering and other abnormal human activities.

Vision-based techniques apart, the de-facto techniques for 3D human pose estimation is optical motion capture (Vicon™[1], MotionAnalysis™[2]) technologies. These technologies require human subjects to wear special suits with optical markers mounted at the specified limb positions. These optical markers are tracked by custom-made cameras made of Infrared(IR) pass-filter coupled with optical lenses. Although these technologies produce highly accurate estimation of 3D human pose, they are expensive, require extensive setup and are intrusive at best.

Even though vision-based system has great potential to be cheap and nonintrusive alternative to optical motion capture system, it remains a challenge to put vision-based system into practical use. State-of-the-art vision-based techniques are still "brittle" compared to the traditional optical motion capture because of lower accuracy and slower speed. It is unfortunate that both criteria are important to meet commercial needs.

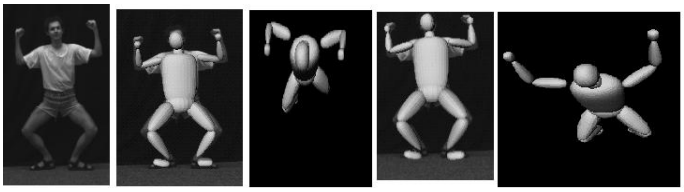

Figure 1 Depth Ambiguities (a,b,c,d,e). (a) Original image (b,d) Different 3D human pose but exhibits similar 2D projection (c,e) Different 3D human pose corresponding to (b,d) respectively. (Figure from [5])

A. Technical Challenges

In general vision based system, 3D pose estimation problem can be addressed by multiple-cameras or single camera. In this paper, we only cover pose estimation using single camera whereas multiple cameras techniques are intentionally omitted because single camera techniques are (1) more challenging and are (2) more practical since most applications only have single camera, for example video surveillance. Interested readers can refer to survey on pose estimation using multiple cameras in [3,4].

According to a survey [5], there are few technical challenges to be addressed in single camera pose estimation such as depth ambiguities (See Figure 1), high-dimensional representation of human pose, self-occlusion, unconstrained motions, observation ambiguities, motion blurs and unconstrained lighting. Depth ambiguities arise because 3D world is projected into a 2D image, causing loss of depth information. Loss of depth information cannot be recovered using single camera only. For this reason, recovering full body human pose is an ill-posed problem since kinematic tree/skeleton is commonly used to represent 3D human body. Without depth information, it is challenging to reconstruct skeleton in 3D. Furthermore, it is common that skeleton is modeled according to human anatomy where limbs (hand, legs, elbow, arms, etc) and torso are modeled using 30-60 joint angle variables. However, estimation of these joint angle variables are computationally expensive because of high dimensional space. Self-occlusion occurs frequently in a single camera view as one body part tends to hide another body part during motion of these articulated limbs. Unconstrained motions are the result of highly diversified human movements, notwithstanding human movements can be highly structured at the same time. For instance, walking or running has repetitive human pose structure over time although it also shows large pose variations at different speed, acceleration and at different body size. Observation ambiguities occurs because single image observation can be mapped to more than one possible 3D human pose, it is

This PhD work is sponsored under Yayasan Khazanah scholarship.

difficult to disambiguate 3D human pose without depth information (See Figure 2). Lighting inevitably changes at different environment as a function of space and time. Lighting variations affect image observations for estimation of body pose. For instance, silhouette shape of the same human pose may appear differently at different capturing time. While capturing rapid human motion, slow camera shutter time causes blurring of image objects, this affects the quality of image observations also.

B. Scope of this Paper

Vision based pose estimation is an intense research topic and has received much attention from computer vision community. This is evident from the substantial number of papers published in this topic found in major computer vision conferences (International Conference on Computer Vision - *ICCV*, Computer Vision and Pattern Recognition - *CVPR*, European Conference on Computer Vision - *ECCV*) and main journals (IEEE Pattern Analysis and Machine Intelligence - *PAMI*, Springer International Journal of Computer Vision - *IJCV*, Elsevier Computer Vision and Image Understanding - *CVIU*) within the last two decades. For example, in a survey done by T. B. Moeslund [3], there are 352 papers (2000 - 2006) in the area of vision-based analysis of human motion. In his survey, research papers are structured in taxonomy according to the sub-problems solved: initialization, tracking, pose estimation and recognition. Of the 352 papers, 125 papers belong to the pose estimation category (2D and 3D inclusive).

Since it is impossible to cover every taxonomy within vision-based human motion analysis, this paper concentrates in the area of 3D human pose estimation with emphasis on single camera view. The main objective of this paper is to review advances made in the single camera 3D human pose estimation within the last ten years. Our goal is to give an overview of the approaches used to solve 3D pose estimation problems by highlighting the mechanism behind such approaches. To this end, Section II shall describes about surveys that have been conducted and published up to this point of writing. Section III shall describes common low-level image features used. Section IV describes types of 3D human body model employed. Section V describes state-of-the-art techniques employed in 3D human pose estimation. Section VI concludes this paper with discussion.

II. Related Surveys

There are few related surveys published in the area of vision-based human motion analysis, all [3,4,6] except one [5] give broad overview of vision-based human motion analysis using the taxonomy of detection, tracking, pose estimation (2D and 3D) and recognition. T. B. Moeslund [3] review advances made in human motion analysis and capture from 2000-2006, extending his previous survey [7] to include new research directions such as detection and tracking of human in natural environment rather than laboratory environment, model-based pose estimation approaches where motion and stochastic sampling framework are employed to search for optimal state (human pose) given the image observations. Their survey divided

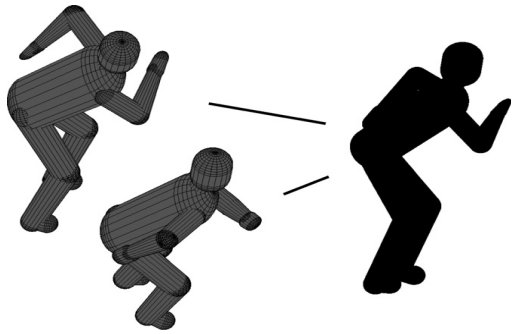

Figure 2 Single image observation (silhouette) can be mapped to two possible 3D human pose. (Figure from [4])

papers into different taxonomy such as initialization, tracking, pose estimation and recognition. R. Poppe [4] review about advances human motion analysis such that human pose estimation problems are divided into two main classes: model-based approaches and model-less approaches. Sminchisescu [5] categorised single camera reconstruction of full-body 3D human motion problems into generative methods or discriminative approaches. Generative approaches build and optimise objective function to match image observations so that the correct human pose hypotheses should maximise observation likelihood within the probabilistic framework. Discriminative methods formulate pose estimation into recognition problem, pose estimation is predicted by trained model using training sets consisting of joint pose and image observations. Discriminative approaches use machine learning extensively to predict state distributions in the absence of depth information. D. A. Forsyth *et al*. [6] review methods to track human body from video. Their review focus on tracking and motion synthesis. 3D body pose can be inferred by lifting 2D pose to 3D pose. They believe that ambiguities during lifting can be partially if not completely solved whenever motion, geometric and context information are incorporated appropriately into probabilistic framework.

Recent new research directions have emerged such as (1) Bottom up approach, detection of local parts (hand, legs, torso) using data-driven approach before pose estimation [8,9,10,11], (2) Learning in low-dimensional *pose manifold* rather than original high-dimensional *appearance manifold*, [12,13,14], and (3) Learning of nonlinear dynamics of human motion models for smoother 3D human pose from video sequence [14,15,16]. The mechanism of these new techniques shall be described in Section V.

III. Low-level Image Observations

Prior to 3D human pose estimation, low-level image observations must be extracted from images or video frames. Ideally, the extracted image observations should encode salient information for subsequent use in high-level image understanding tasks. In this context, high-level understanding tasks here refer to 3D body pose estimation. Although in theory, original images or video sequences could be used as image observations, they are too "noisy" to be of any use for high level understanding tasks. Such noisy information originated from lighting variations, different clothing, cluttered background can seriously undermine high level understanding tasks. Therefore, low-level image

observations to be extracted during feature extraction process is highly task specific. It is common computer vision practice to make assumptions about the environment, acquisition, image generation process in order to simplify feature extraction process. In statistical perspectives, image observations extracted must have strong correlation to the problems at hand so that variations of estimation can be minimised. Often in 3D body pose estimation problem, commonly used image observations are silhouette and shapes, edges, motions, colours and recent approaches are combinations of them.

A. Silhouette and Contours

From visual observation, silhouette shows strong correlation with body contours. Different human pose exhibits different silhouette shape. Therefore it is safe to assume that there exists functional mapping between silhouette and human pose. Besides, there exists methods that can reliably extract silhouettes from video sequences considering background is fairly static. Under stable background lighting condition, background subtraction with single Gaussian distribution on colour statistics is deemed sufficient [17]. Silhouette is insensitive to colour and textures while irrelevant to clothing types. Agarwal [18] extracted silhouette using background subtraction, and then shape contexts [19] (image observations) are extracted from the sampled silhouette edges to form histogram features (See Figure 3). In certain cases, feature extraction techniques must be robust in the presence of noisy silhouette such as in [12]. This scenario occurs frequently in natural images where scenes are cluttered with other irrelevant objects. One limitation with silhouette is that when using single camera, depth information is lost hence 3D pose cannot be reliably recovered due to observation ambiguities.

B. Edges

Edges information can be robustly extracted at low computation costs. This information can be used as boundary to delimit body parts, for instance from the observation clear outline can be found between arms and torso boundary. Edges are also invariant to colour, texture and lighting variations. Deutscher [20] use gradient-based edge detection mask to detect edges, threshold is then applied to eliminate noisy edges to produce pixel map and later used in the weighting function. However, edges information are not robust against cluttered background, many false positives may result as shown in [21]. Post-processing is typically required to achieve reliable detection whenever background is cluttered. Ramanan *et al* [21] enforce constant appearance and integrate appearance information besides edges to reliably find body segments. Their method demonstrates good performance in 2D tracking of body parts in low-resolution natural video, frequent inter-body occlusion and cluttered background even works for long video sequence. Wu [11] trained human detectors by combining body part detectors (head-shoulder, torso and legs) in Adaboost framework using edgelet features. Edgelet features are

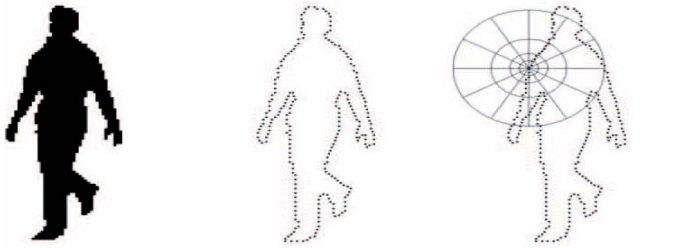

Figure 3 (a) Silhouette extracted by background subtraction (b) Sampled edge points. (c)Shape contexts computed on edge points. (Figure from [18]).

essentially short segment edges computed using Sobel masks quantised into six discrete orientations.

C. Motions

Motions information extracted from video sequences are used frequently in human pose tracking [13], motion segmentation and object segmentation. Motion can be measured using optical flow methods which can be categorised into dense or sparse methods. In general, dense optical flow is an image map where each pixels in the image map encodes 2D velocity indicating magnitude of 2D displacement of a pixel for two subsequent frame. Readers can refer to survey by Beauchemin [22] for the state-of-the-art techniques in computation of optical flow. Urtasun [13] use optical flow as an additional source of information besides silhouette to compute 2D point correspondence in pair of two consecutive frames. This 2D point correspondence is part of the objective function to be minimised to obtain smoothness in the time varying 3D human pose in video.

D. Colours

Colour features can be important cues for the positions of different body parts for example face, arm, and legs. Strength of colour features lies in the fact that these features are invariant to scale and body poses, robust under stable lighting conditions and require low cost computation. In most cases, colour features are effective measures for skin colour detection algorithm where pixels are evaluated by learnt skin colour histograms [8] model. Lee [24] initialise body part detection using skin colour features to find coarse positions of face, arms, and sometimes legs. In his work, an image is segmented into regions using colour-based segmentation. For each segmented region, probability is computed by evaluating the region against skin-colour histograms. Extracted elliptical regions with high skin probabilities are used to predict the positions of heads and limbs. One disadvantage of using colour-based features are many false positives or negatives expected in situation where non-skin objects have similar skin colour characteristics while skin objects are occluded by non-skin colour objects. Post-processing steps are typically required to ensure consistent detection. For instance, skin colour used with geometrical constraints are used to locate candidate face locations given that only good elliptical region are considered.

E. *Combination of Image Observations*

Intuitively, if multiple image observations/cues are combined, they should provide better performance than single image observation. The reason for this is because one image observations may failed in certain scenario but fare better on others. One simple strategy for combination is to use weighted image observations, such that weights are set heuristically or learned using data-driven approach. In [20], objective function is constructed using both silhouette and edges features later used by stochastic sampling framework to track 3D human pose from video sequences. Lee [24] combine colour, shape and contour features in weighting methods to find candidate head location. Ramanan [8] use edges and colour information. Sminchisescu [24] combine silhouettes (by extracting shape contexts features [19]) and internal edges into a single image vector before clustered by K-means. However, care must be taken to build observation likelihoods in such a way that good image observations are given more weights than the bad image observations to achieve better performance.

IV. 3D HUMAN BODY MODEL

Full human body is highly articulated structure but body parts can be considered as rigid structure. To that end, there are few body models to represent articulated 3D human pose. In most cases, 3D human pose can be represented by kinematic tree model (see Figure 4(a)), consisting of segments linked by joints. In kinematic tree model, joints are consider nonarticulated and can have maximum three degrees of freedom (DOF) corresponding to three orthogonal directions. Number of joints and DOF required depends on the degree of details required in the application. Whereas for spatial resolution of human in the image, smaller the spatial resolution of the human, smaller the number DOF required and vice versa. Even though detailed DOF can produce more realistic human pose, it also increases computational complexity as estimation now has to be performed at higher-dimensional space. Therefore trade-off must be made to balance degree of details required against computational complexity. Papers that use kinematic tree models are [12,13,18,20,23]. Besides kinematic model, human pose can also be represented by volumetric models such as elliptical cylinder (See Figure 4(c)). This volumetric model have all the limbs fleshed out by elliptical cylinder, papers using this model are [20,25,23]. Lee [23] represents 3D human body by both kinematic tree model and volumetric elliptical cylinder. His model can simultaneously describes human shape of different body size and clothing that a person wears. Another volumetric models are super quadrics [26] and generalised cones [27]. Volumetric base models representation have limited description capabilites when comes to variations of body size. In most older works, width or length of limbs in elliptical cylinder model is manually fixed during initialisation for computational convenience. However, some researchers [23,28] have started to take this issue into consideration by recovering the parameters of limbs *automatically* during initialisation. Of course, this will takes additional steps thereby increasing computational complexity.

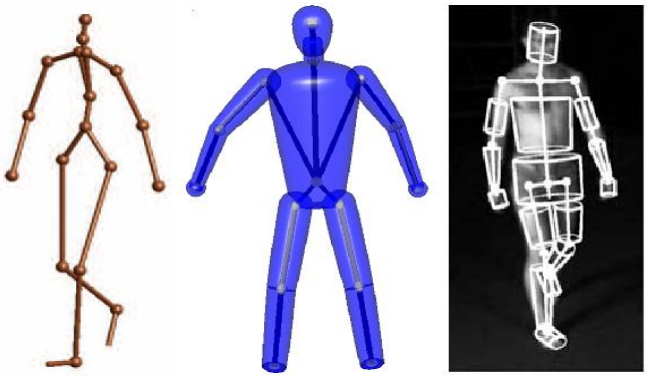

Figure 4 Human body model to represent 3D human pose (a) Kinematic model characterised by segments linked together by joints (Figure from [18]) (b) 3D volumetric model consisting of super quadrics (Figure from [26]) (c) Cylinder with elliptical cross-sections representing each limbs (Figure from [20]).

V. SINGLE CAMERA 3D HUMAN POSE ESTIMATION

Pose estimation can be formulated as optimisation problem since the main concern is to find the pose parameters that minimise errors between the image observations and the 2D projected 3D body pose. Human pose estimation can also be regarded as inferring the underlying kinematic structure (in the form human skeletal) from image observations. In single camera problem, it is common to impose additional constraints (kinematics, motions, or geometrical information) to reduce the ambiguities that appear in the inferred 3D body configurations. Without use of constraints, there are many possible 3D configurations that could explain a single observation manifested in multiple modes in likelihood function.

To deal with multiple modes (peaks) problem in the posterior (likelihood function), randomised search in the form of stochastic sampling method (*particle filters*) are commonly employed. Few publications explaining the idea behind particle filters such are given in [29,30,20].

The key idea behind particle filters is to use randomised sampling to search the posterior because the distribution of posterior is non-Gaussian in general. To that end, particle filters approximate the posterior distribution by sets of points concentrating around places where large values of likelihood are found. The posterior evolves over time using assumed underlying dynamical state space model to predict time-varying configurations. The sampled representation of the new posterior are then the predicted *prior* of body configuration. The prior body configuration is later matched to image observations, and those sets of points that give good comparisons are given more weights, resulting in new representation of the desired posterior distribution. Particle filters are becoming important for applications that need approximate model to explain the underlying dynamics of time-varying physical system that typically shows nonlinearity/non-Gaussianity (such as when Kalman filtering fails) in the their posterior distribution. For instance, one can use particle filtering to track multivariate data in time-series problem. A good tutorial of using particle filters on tracking problem can be found in [31].

However, one disadvantage of using particle filters in estimating 3D body pose is high data dimensionality. One 3D human body can have at least 20 degrees of freedoms according to D. Forsyth [6] (one at each knee, two at each hip, three at each shoulder, one at each elbow and six for the root). To deal with problems of high-dimensionality, researchers focus their efforts on building more efficient search methods rather than improving the core algorithm of particle filters.

Sidenbladh *et al* [25,32] use *importance sampling* method to guide particle filtering search on likelihood either on the learnt walking model or on motion database. Importance sampling method use proposal distribution as alternative to the prior likelihood function so that samples can be drawn in places that are more likely. Another variant of importance sampling is to use annealed search by Deutscher *et al* [20]. This method shows improvement in speed than the former method in tracking 3D pose of a subject. However, this method uses three cameras to track a single person under simple black background. Moreover the method shows no experimental evidence that it can work in natural environment when clutters and other textures are all too common.

Sminchisescu *et al* [33] present an extension to particle filtering method to recover 3D human pose from single camera image sequences. They believe that for successful single camera 3D body tracking, at least three difficulties need to be resolved such as (1) must be able to estimate 30 joint parameters, (2) estimation of depth information to recover the unobservable 1/3 DOF (3) matching of image observations to complex body model under self-occlusion and cluttered background. They design cost matching function using combination of edge, optical flow and motion boundaries while enforcing hard joint angle limits with non-self-intersection constraints. *Covariance scaled sampling* method is used as search strategies to find good poses in high-dimensional body configuration space. From their observations, cost minima occur most likely along the local valley of cost surface where covariance has highly uncertain directions. The hypothesis distribution is determined along these highly uncertain directions while combining with certain temporal dynamic models. Good 3D human body poses are sampled using random or regular pattern from this hypothesis distribution for rescaled covariances. Results demonstrate robust tracking of entire arm and full-body 3D for those video sequences that contain self-occlusion and cluttered background. In further research [34], they improve the speed of the search for local minimum by constructing an interpretation trees that can generate many possible 3D human poses to explain the same image observations by introducing inverse kinematics. Experiments show that it can track 3D human body in short video sequences that contain fast, unpredictable and complex motion (such as dancing) under cluttered backgrounds.

In recent work, Lee *et al* [23] introduce novel methods to address *automatic* initialisation issue in monocular 3D human pose tracking by estimating multiple persons positions and sizes before inferring their corresponding 3D poses. Automatic initialisation is seldom addressed and is an important step to bootstrap pose estimation for many applications. In their approach, body part positions are estimated by the head, shoulders and limbs detector modules. Multiple image cues/observations such as skin colour, head shoulder shapes are used in tandem with learnt detector to locate the head position. Torso location is estimated based on head position (just below the head). In the second stage, belief propagation technique is used to refine the positions of the body parts from the earlier detection. In the last stage, data-driven Markov chain Monte Carlo (MCMC) [35,36] algorithm is used to estimate 3D human pose in each frame. MCMC is a variant of stochastic sampling method used in particle filtering framework to explore solution spaces by carefully designing proposal function to generate optimal candidate states (body pose). In their work, proposal function to evaluate likelihood of a candidate state is formulated using four criteria (1) region consistency (2) colour dissimilarity with background (3) skin colour and (4) foreground matching.

Although human pose can be estimated from static images only, motion prior helps in smooth tracking of human pose by enforcing strong constraints. Recent approaches mainly concentrate on learnt motion model to obtain realizable human poses and motions [32,37,38]. Learnt motion models often use motion capture training data. One best known publicly available motion capture database is from CMU[39]. Nevertheless, one weakness with learnt motion models approach is the excessive dependency on the amount and quality of motion capture training data. There must be sufficiently large amount of training data to accurately learn the representation of all possible human motions. Besides, some motion capture data is noisy because of the presence of random and systematic errors inherent during human motion measurement capture.

Howe *et al* [38] present a system to reconstruct 3D motion of human using single camera. In each frame, they track the entire body using learnt 2D body parts detector. Motion capture data is assembled into *snippets* (motions consisting of 11 successive frames) are later used to train mixture-of-Gaussians probability density functions (pdf) for few classes of 3D body configurations. The output of 2D body parts are matched to the pdfs to find the corresponding 3D body configuration. Siddenbladh *et al* [32] solve similar problem using *generative model*. Generative model determines the likelihood of observing certain image observations (shape, appearance and motion features) given a state (3D human pose and movement). Learnt motion models are determined using previous history of states from motion capture data. Particle filtering optimisation is later used to find the approximate states. Agarwal *et al* [40] present a novel approach that is able to track *unseen* human pose - not in motion capture training data - in the presence of complex background. Their method track 2D human pose but readily extends to 3D when needed. Rather than learning the whole state space parameters, they partition the state space parameters into regions with similar dynamical characteristics. This facilitates learning of nonlinear dynamics using piece-wise linear autoregressive process for each region. Bottom-up processes are also been incorporated

into top-down processes as illustrated in [41]. The advantages of this approach are *automatic initialisation* in contrast to manual initialisation and recovery from tracking failures. In [41], body is represented as graphical model. Each nodes corresponds to a body part and edges between nodes represents statistical dependencies and physical constraints. Probabilistic models are learned from motion capture training sets to capture temporal evolution of each nodes over time. 3D human pose at any time instant is then recovered by probabilistic inference using non-parametric belief-propagation [42].

One major problem in estimating 3D human pose is high dimensionality of pose space data. This problem poses great challenge to machine learning approaches including particle filtering framework because number of training samples available are often too small to cover all possible human movements intricacies. This phenomenon is the well known "curse of dimensionality". To avoid this phenomenon, recent research trends concentrate on discovering methods to reduce the data dimensionality prior to estimation [37,43]. Moreover, these approaches are also motivated by the findings that human pose can be sufficiently represented by low-dimensional latent *manifolds* as shown in [37]. To that end, these approaches attempt to learn low-dimensional mapping functions relating human pose to the image observations. However, learning mapping functions are difficult because the manifolds are nonlinear. To recover 3D human pose from image observations, two mapping functions are learned such as mapping of image observations space to pose space and its corresponding inverse mapping.

Elgammal *et al* [37] introduce method to reconstruct 3D body from a given viewpoint from silhouettes information using single camera. Local linear embedding (LLE) [44] are used to learn mapping of pose space to silhouette space and its corresponding inverse mapping. Unseen 3D human pose - not part of the training data - is recovered by interpolation by radial-basis function (RBF). While in their earlier work [37] is strictly view dependent and limited to walking pose activity, Elgammal *et al* [16] in his later work model 3D body pose of a person observed at different viewpoints and extensible to general human motions. Body pose and viewpoint are explicitly modeled in two separated low dimensional representations. In similar spirit, Sminchisescu *et al* [43] propose the use of spectral embedding [46] algorithm to learn mapping of image observation space to low-dimensional manifold pose space and its inverse mapping separately. Tracking of human pose is later constrained to the learnt low-dimensional manifolds. Agarwal *et al* [18] *implicitly* achieve data dimensionality reduction using relevance vector machines (RVM) regression. RVM selects only the "most relevant" basis function by retaining only the relevant input features. As a consequence, large training data are reduced to a minimal subset.

Although LLE and spectral embedding methods can learn low-dimensional embedding manifolds from data, they lack probabilistic interpretation. This suggests that no straightforward learning-based method can be applied to the learnt low-dimensional manifolds. It is also difficult if not impossible to find inverse mapping between low-dimensional latent space back to image observations space. Urtasun *et al* [47] propose the use of Scaled Gaussian Process Latent Variable Model (SGPLVM) [47] that admits probabilistic interpretation to learn human pose prior with continuous mapping between observation space and pose space. Human pose is recovered by finding body pose that maximise the likelihood of the learnt SGPLVM model given sets of image observations. Results demonstrate good tracking accuracy of 3D body pose for both walking and golfing activity but in their experiments 3D body positions are manually initialised. SGPLVM has generalise well even with small training set available. In similar work, SGPLVM was extended by incorporating additional nonlinear dynamics mapping while retaining the original mapping and its corresponding inverse mapping obtained through Gaussian Process Dynamical Model (GPDM) [14]. GPDM produces smoother motion models compared to SGPLVM. In recent work, the fact that the tasks of body estimation is strongly correlated with activity recognition motivated T. Jaeggli *et al* [12] to introduce method to simultaneously track human pose and recognition of multiple action categories. In similar spirit they use LLE as mapping of body pose to low-dimensional space, while kernel regressor as inverse mapping back to original body pose. Low-dimensional models are learned separately for different activities for instance each low-dimensional manifolds are dedicated to walking and running respectively. To model activity switching, nonlinear mapping between pair of activities are also modeled. Likelihood function using gaussian distribution are used as probability measures to determine activity transition. Experiments demonstrate that their approach can reliably track subjects while recognising activity transition simultaneously even with low-resolution video.

Vast pose estimation literature sees major problem lies in managing multiple modes in likelihood function in high-dimensional data, which explains the extensive use of particle filtering variants. Some evidence [6], though inconclusive, seems to suggest that ambiguities may not persist when short motions (*snippets*) are used in place of single frame. Howe *et al* [49] reconstruct 3D body pose by comparing 3D motion capture data with 2D snippets via dynamic programming. Ramanan *et al* [50] use similar approach to lift 2D snippets into full 3D body pose by matching them to the stored 3D motion capture database with assumption camera is in lateral view. Best matching 3D pose is also recovered by dynamic programming. The same approach is being used in [51] to build viewpoint invariant human activities retrieval system. 3D body pose is constructed by lifting the output from 2D limb detectors [9]. One disadvantage of this approach is one needs large 3D motion capture database in order to lift 2D pose into good 3D pose let alone massive computation incurred during the matching process. Therefore, this approach remains as an open research problem. It remains to be seen on how one can reconstruct 3D body pose from 2D snippets using smaller motion capture database.

VI. DISCUSSION

Single camera 3D human body estimation is an active area of research as evident in the substantial publication of 3D body pose estimation techniques within the last two decades. This is mainly motivated by the facts that vision-based system is far more cheaper and non-intrusive compared to the conventional optical motion capture technologies [1,2]. However, single camera pose estimation is challenging because the presence of large variations in human motion and appearance, different camera viewpoint, high-dimensional data and changing environment. Though challenges above are generic in computer vision, visual ambiguity inherent in single camera are the main issues. Visual ambiguity is direct result of suppressed depth information, self-occlusion, unconstrained general motions and observation ambiguities.

Most of the literature reviewed handle visual ambiguities using particle filtering variants [5,20,25,30,32,33,34,36]. One possible explanation of why particle filtering method is popular is because of the earlier success of particle filtering in visual tracking under cluttered environment [30]. It is not unusual to see different search techniques are proposed to guide particle filtering method to recover optimal body pose [20,33,34]. For instance joint angle limitations constraints and other contexts (skin colour, kinematic constraints) are used as search technique to prune improbable solution space. Pruning technique is the key point to successfully disambiguate visual ambiguities that manifest themselves as multimodal posteriors.

On the other hand, significant advances are made to reduce high-dimensionality of pose space since there is evidence showing 3D human pose often falls into low-dimensional manifolds [37]. This serves as the main motivation for other similar works [12,16,37]. However, once pose space is mapped to low-dimensional manifolds, the corresponding inverse mapping must also be defined so that the original pose space can be recovered for evaluation of likelihood function. This task is challenging since low-dimensional manifolds are in general nonlinear. Furthermore, most work only demonstrates narrow classes of human pose (walking and running) under a given viewpoint. It remains to be seen whether further research can advances these techniques into general human movements with different viewpoints. Recent work to address these issues are by Elgammal *et al* in [16].

Tracking of human pose in 3D can be considered as time-series modeling since body pose evolves over time. Each body pose represents sampled output state from the system at a particular time, *t*. Thus, it is intuitive to incorporate motion prior for smoother estimation of human pose in tracking framework. In most literature, learnt motion model is obtained by tracking in the learnt low-dimensional manifolds [13,14,15]. Recent work [12] also shows that learning in low-dimensional manifold have strong correlation with recognition of human activity.

In future work, there are few promising research directions to be explored. One is to design good features. It is surprising that not many literature explore methods to construct good features. The point is intuitive: Good features when combined with simple inference is better than using bad features combined with sophisticated inference method. One good example done in this aspect is by Mori *et al* [52,53]. It seems in their work visual ambiguity are reduced. Also, further research can be done to solve *automatic* initialisation and recovery of human pose in 2D before lifting to 3D body pose. 2D pose estimation is more effective than 3D in solving self-occlusion, a much ignored problem. Furthermore, recent works have seen progresses made in development of robust 2D part detectors [8,9,10,11,21]. One can take advantage of 3D motion capture database to be used for matching output from 2D part detectors to recover unambiguous 3D body pose. Here, the open research problem is on how to combine 2D part detectors with 3D pose estimation into one framework. From practical perspective, methods must be aimed at reducing algorithm computational complexities. Finally, to evaluate the effectiveness of pose estimation methods, common database known as HumanEva-I database [54] could to be used since evaluation criteria are generally accepted.

ACKNOWLEDGMENT

The authors would like to thank anonymous reviewers for their comments on the manuscript. This PhD work is fully funded by the Yayasan Khazanah scholarship- Khazanah Nasional.

REFERENCES

[1] Vicon Optical Motion Capture System. Vicon Motion System Ltd., Oxford, UK. http://www.vicon.com.

[2] MotionAnalysis™ Motion Capture System. MotionAnalysis Corporation. Santa Rosa, CA USA. http://www.motionanalysis.com.

[3] T.B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding(CVIU),* vol. 104 (2-3), pp. 90-126, 2006.

[4] R. Poppe, "Vision-based human motion analysis: An overview," *Computer Vision and Image Understanding(CVIU),* vol. 108, pp. 4-18, 2007.

[5] C. Sminchisescu, "3D Human Motion Analysis in Monocular Video: Techniques and Challenges," in *Human Motion - Understanding, Modelling, Capture and Animation,* vol. 36, A. Elgammal, B. Rosenhahn, K. Reinhard, Eds. New York: Springer, 2008, pp. 185-211.

[6] D. A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan, "Computational Studies of Human Motion: Part 1, Tracking and Motion Synthesis," in *Foundations and Trends® in Computer Graphics and Vision,* vol. 1 (2-3), 2006, pp. 77-254.

[7] T.B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer Vision and Image Understanding(CVIU),* vol. 81 (3), pp. 231-268, 2001.

[8] D. Ramanan and D. A. Forsyth, "Finding and Tracking People from Bottom Up," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2003, pp. 467-474.

[9] D. Ramanan, D. A. Forsyth and A. Zisserman, "Strike a Pose: Tracking people by Finding Stylized Poses," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2005, pp. 271-278.

[10] B. Wu and R. Nevatia, "Tracking of Multiple, partially occluded humans based on static body part detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2006, pp. 951-958.

[11] B. Wu and R. Nevatia, "Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors," *International Journal of Computer Vision(IJCV),* vol. 75(2), pp. 247-266, 2007.

[12] T. Jaeggli, E.K. Meier and L. V. Gool, "Learning Generative Models for Multi-Activity Body Pose Estimation," *International Journal of Computer Vision(IJCV),* vol. 83(2), pp. 121-134, 2009.

[13] R. Urtasun, D. J. Fleet and P. Fua, "Temporal motion models for monocular and multiview 3D human body tracking," *Computer Vision and Image Understanding(CVIU),* vol. 104, pp. 157-177, 2006.

[14] R. Urtasun, D. J. Fleet and P. Fua, "3D People Tracking with Gaussian Process Dynamical Models," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2006, pp. 238-245.

[15] J. M. Wang, D. J. Fleet and A. Hertzmann, "Gaussian process dynamical models," in *Neural Information Processing Systems,* 2006, pp. 1441-1448.

[16] C. S. Lee and A. Elgammal, "Modeling view and posture manifolds for tracking," *International Conf. on. Computer Vision,* 2007, pp. 1-8.

[17] C. Stauffer and W. Grimson, "Adaptive Background Mixture Models for Real-time Tracking," in *Proc IEEE Conf. Computer Vision and Pattern Recognition*, 1999, pp. 23-25.

[18] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *IEEE Transactions on Pattern Analysis and Pattern Recognition(PAMI)* vol. 28, pp. 44-58, 2006.

[19] S. Belongie, J. Malik and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," *IEEE Transactions on Pattern Analysis and Pattern Recognition(PAMI)* vol. 24(4), pp. 509-522, 2002.

[20] J. Deutscher, and I. Reid, "Articulated Body Motion Capture by Stochastic Search," *International Journal of Computer Vision(IJCV),* vol. 61(2), pp. 185-205, 2005.

[21] D. Ramanan, D. A. Forsyth, and A. Zisserman, "Tracking People by Learning Their Appearance," *IEEE Transactions on Pattern Analysis and Pattern Recognition(PAMI)*, vol. 29, pp. 65-81, 2007.

[22] S. S. Beauchemin and J. L. Barron, "The Computation of Optical Flow," *ACM Computing Survey,* vol. 27(3), pp. 433-466, 1995.

[23] M. W. Lee and R. Nevatia, "Human Pose Tracking in Monocular Sequence using Multilevel Structured Models," *IEEE Transactions on Pattern Analysis and Pattern Recognition(PAMI)*, vol. 31(1), pp. 27-38, 2009.

[24] C. Sminchisescu, A. Kanaujia and D. N. Metaxas, "BMA³E : Discriminative Density Propagation for Visual Tracking," *IEEE Transactions on Pattern Analysis and Pattern Recognition(PAMI)*, vol. 29(11), pp. 2030-2044, 2007.

[25] H. Siddenbladh, M.J. Black, L. Sigal, "Implicit Probabilistic models of human motion for synthesis and tracking," in *Proc. European Conf. on Computer Vision,* 2000, pp. 784-800.

[26] R. Kehl and L. V. Gool, "Markerless tracking of complex human motions from multiple views," *Computer Vision and Image Understanding(CVIU),* vol. 104(2-3), pp. 190-209, 2006.

[27] D. Gavrila, "Vision-based 3D Tracking of Humans in Actions," PhD thesis, *Department of Computer Science,* University of Maryland, 1996.

[28] J. Carranza, C. Theobalt, M. A. Magnor and H-P. Seidel, "Free-viewpoint video of human actors," *ACM Transactions on Computer Graphics*, vol. 22(3), pp. 569-577, 2003.

[29] A. Doucet, N. D. Freitas and N. Gordon, *Sequential Monte Carlo Methods in Practice.* Springer-Verlag, 2001.

[30] M. Isard and A. Blake, "CONDENSATION - Conditional density propagation for visual tracking," *International Journal of Computer Vision(IJCV),* vol. 29(2), pp. 5-28, 1998.

[31] M. S. Arulampalam, S. Maskell, N. Gordon and T. Clapp, "A tutorial on particle filtering for online nonlinear/non-Gaussian bayesian tracking," *IEEE Transactions on Signal Processing,* vol. 50(2), pp. 174-188, 2002.

[32] H. Sidenbladh, M. Black and D. Fleet, "Stochastic tracking of 3D human figure using 2D image motion," in *Proc. European Conf. on Computer Vision,* 2000, pp. 702-718.

[33] C. Sminchisescu and B. Triggs, "Covariance scaled sampling for monocular 3D body tracking," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2001, pp. 447-454.

[34] C. Sminchisescu and B. Triggs, "Kinematic jump processes for monocular 3D human tracking," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2003, pp. 69-76.

[35] Z. W. Tu, S. C. Zhu and H. Y. Shum, "Image segmentation by data driven markov chain monte carlo," in *Proc IEEE International Conf. on Computer Vision*, 2001, pp. 131-138.

[36] M. Lee and I. Cohen, "Proposal maps driven MCMC for estimating human body pose in static images," in *Proc IEEE Conf. on Computer Vision and Pattern Recognition*, 2004, pp. 334-341.

[37] A. Elgammal and C. S. Lee, "Inferring 3D body pose from silhouettes using activity manifold learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2004, pp. 681-688.

[38] N. R. Howe, M. E. Leventon and W. T. Freeman, "Bayesian reconstruction of 3D human motion from single-camera video," in *Neural Information Processing Systems,* 2000, pp. 800-826.

[39] CMU Graphics lab motion capture database, Carnegie Mellon University, http://mocap.cs.cmu.edu.

[40] A. Agarwal and B. Triggs, "Tracking articulated motion using a mixture of autoregressive models," in *Proc. European Conf. on Computer Vision,* 2004, pp. 54-65.

[41] L. Sigal, S. Bhatia, S. Roth, M. J. Black and M. Isard, "Tracking loose-limb people," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2004, pp. 421-428.

[42] E. B. Sudderth, A. T. Ihler and W. T. Freeman, "Nonparametric belief propagation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2003, pp. 605-615.

[43] C. Sminchisescu and A. Jepson, "Generative modeling for continuous non-linearly embedded visual inference," in *Proc. ACM International Conf. on Machine Learning,* 2004, pp. 759-766.

[44] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290(5500), pp. 2323-2326, 2000.

[45] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15(6), pp. 1373-1396, 2003.

[46] R. Urtasun, D. J. Fleet, A. Hertzmann and P. Fua, "Priors for people tracking from small training sets," in *Proc IEEE International Conf. on Computer Vision*, 2005, pp. 403-410.

[47] N. D. Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data," in *Neural Information Processing Systems,* 2004, pp. 329-336.

[48] K. Grochow, S. L. Martin and A. Hertzmann, "Style-based inverse kinematics," in *Proc SIGGRAPH,* 2004, pp. 522-531.

[49] N. R. Howe, Silhouette lookup for automatic pose tracking," in *IEEE Workshop on Articulated and Non-rigid Motion,* 2004, pp. 15-22.

[50] D. Ramanan and D. A. Forsyth, "Automatic annotation of everyday movements," in *Neural Information Processing Systems,* 2003, pp. 329-336.

[51] N. Ikizler and D. A. Forsyth, "Searching for complex human activities with no visual examples," *International Journal of Computer Vision(IJCV),* vol. 80(3), pp. 337-357, 2008.

[52] G. Mori and J. Malik, "Recovering 3D human body configurations using shape contexts," *IEEE Transactions on Pattern Analysis and Pattern Recognition(PAMI)*, vol. 28(7), pp. 1052-1062, 2006.

[53] G. Mori and J. Malik, "Estimating human body configurations using shape context matching," in *Proc. European Conf. on Computer Vision,* 2002, pp. 666-680.

[54] L. Sigal, M. 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, 2006.