Single-view and Multi-view Methods in Marker-less 3D Human Motion Capture

Tong Xu
School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing, Jiangsu province, China
1171337410@qq.com

Abstract. Human motion capture has now played a pivotal role in more and more applications, including biomechanics, sports, image segment, animation, robotics, etc. Although commercial marker-based human motion capture models have matured, the shortcomings, such as obtrusion, expense, errors due to damage to the marker trajectories, long set-up times and etc. exposed by this approach are becoming more and more apparent. Marker-less human motion capture analysis is likely to provide inexpensive and efficient solutions to solving these problems for the reconstruction of human motion in the future. In this paper, we discuss and compare the background and characteristics of marker-based and marker-less human motion capture models. Then we divide the marker-less human motion capture into single view and multi view and display some popular models. These methods are also categorized according to their internal logic and algorithms. Finally, we present some of the major shortcomings of the current marker-less human motion capture models and the future direction of development.

1. Introduction

Nowadays, with the development of computer and Artificial Intelligence, the analysis of human actions by a computer is getting more and more mature. Also, motion capture (MoCap) is the basic technology for many applications in the field of computer vision and computer graphics. From 1980s to 2000, Moeslund and Granum [1] show that human motion capture covers many aspects, capturing and connecting large-scale body movements, like the motions of head, arm, torso and legs. Then, many researchers have conceived a variety of methods for studying human motion capture. The first and most popular method is that markers are attached to human body. However, marker-based human motion capture suffers from well-known disadvantages such as obtrusion, expense, errors due to damage to the marker trajectories, long set-up times, and the incompetence to simultaneously capture the dynamic shape and motion of actors in normal clothing. Therefore, the marker-less motion capture models are under studied, which can overcome some of these problems have emerged. Compared to commonly used marker-based approaches, marker-less methods are based on sensor data (usually images) without special preparation of the subject. Elhayek et al. [2] show that marker-less skeletal motion tracking is achievable and performs well in less controlled studio settings or outdoors, as well as in front of more general backgrounds where foreground segmentation is hard. In the field of biomechanical applications, Ceseracciu et al. [3] show that although in the past decade, marker-less motion capture technology has gained more and more attention in the field of biomechanics. In the clinical field, because there are few papers to compare the difference between the marker-based method and the existing method, there is some controversy about the marker-less approach. Therefore,
Ceseracciu et al. study the application of marker-less technology in data obtained with marker-based systems. Zhou et al. [4] show that due to the increase in the number of patients with stroke or other motor dysfunction, the related issues of human motion tracking and rehabilitation have become more and more active. Zhou et al. describe recent advances in human motion detection/tracking systems, particularly existing or potential applications of stroke rehabilitation and summarize the major achievements of related systems. Except for biomechanics, the human motion capture model is used in many other applications. Shingade et al. [5] tell us about the feasibility of using the human motion capture models applied to sports. Liu et al. [6] can use the human motion capture in the field of image segment. Sturman et al. [7] make a brief description of motion capture for computer character animation. Field et al. [8] combine the human motion capture sensors and analysis with the robotics.

This paper is going to discuss the marker-less motion capture from the perspective of its categories, single-view MoCap and multi-view MoCap, including commonly used algorithms and recent applications. To be specific, we will firstly start from these two aspects to introduce which areas of single view MoCap and multi-view MoCap are applicable, and the areas that have not been accessible so far. Then, through the comparisons, this paper will also suggest other ways to further improve the marker-less motion capture models. Finally, this paper will give the applications and contributions of today's marker-less motion capture models in various fields.

2. The overview of human motion capture: marker-based vs marker-less
Since the 1870s, some scientists around the world have started to study the human motion capture. The research processes have not been smooth and conclusions have undergone various evolutions.

Moeslund and Granum [1] narrate that from 1980 into the first half of 2000, various categories may be used for a taxonomy. Fleisig, Barrentine and Zheng [9] introduce the kinematic and kinetic comparison of baseball pitching among various levels of development. Also, Riley, Paolini and Croce [10] deal with the comparison of kinematic and kinetic in human motion capture (kinetic vs kinematic). Gavrilla and Davis [11] mention 3D model-based tracking of humans in action and Moses [12] talks about the limitations about non-model-based recognition schemes which refers to approaches to visual object recognition (model-based vs non-model-based). Mian, Bennamoun and Owens [13] narrate a 2D-3D hybrid method to automatic face recognition (2D approaches vs 3D approaches). Shafer, Stentz and Thorpe [14] mention an architecture for sensor fusion in a mobile robot (sensor modality, visible light, infrared (IR) light, range data, etc.). Tsitsiklis [15] talks about the utilization of a large number of different sensors. Cheu, Lee and Xie [16] analyze the data from mobile and stationary sensors (mobile vs stationary sensors). Gray, Brennan and Tao [17] mention that recognition, reacquisition and tracking are three of the most important topics in surveillance research (tracking vs recognition). Mykhaylo, Stefan et al. [18] are committed to solve automatic recovery of 3D human pose from monocular image sequences (pose estimation vs tracking). Aldoma, Marton and Tombari [19] deal with 3D object recognition and 6 DOF pose estimation (pose estimation vs recognition). Cao, Simon and Wei [20] talk about multi-person 2D pose estimation using part affinity fields related to human motion capture (one person vs multiple persons). Also, there are some other classes like distributed vs centralized processing, various motion-type assumptions (rigid, non-rigid, and elastic), etc.

Researchers have chosen different categories for corresponding research, using different algorithms and assumptions to make significant contributions to the perfection of human motion capture technology. In the course of the research, many problems are also discovered. For instance, many motions become blurred when projected into an image plane; movements along the optical axis are difficult to record robustly; a lot of time is spent in recording and marking training data and the like. Moreover, Moeslund, Hilton and Volker [21] review the trend of video-based human motion capture from 2000 to 2006 and discuss the open problems of future research to enable automatic visual analysis of human motion. And during this time period, the human motion analysis system, namely human detection, tracking and activity comprehension, has been further developed. Since 2006, with the improvement of the quality of hardware and software equipment, a large number of researchers
have focused on 3D and modeling technology. For instance, Li, Yang and Sclaroff [22] use a coordinated hybrid monocular technique of the factor analyzer to track 3D human motion. Therefore, the first and most popular method is attaching markers to human body. However, people have studied this section and found that this model has obvious drawbacks, like obstruction, expense and long setup time. Thus, the marker-less motion capture models that can overcome some of these problems have emerged. In the kinematic skeleton model, as long as there are enough cameras (more than 7), efficient high-precision marker-less tracking can be achieved using the local pose optimizers.

Besides, with the continuous improvements of human motion capture technology in the past 100 years, some of its applications have also developed rapidly, such as the quality of the sensor, the speed of data processing, etc. Field et al. [23] show that robot programming through demonstrations has a relatively long history. In recent years, research on faster programming of industrial robots has been extended to mimic robots by using motion capture technology and machine learning. The methods of human motion tracking in the above-mentioned thesis of robot research paper highlight the potential advantages of each sensing mechanism. Tracking progress in this area is very important, because the technology is new and rapidly changing, especially for current trends in analytical methods.

3. Single-view methods

3.1. Factorizing 2D observations to obtain convincing 3D reconstructions

When studying the motion problem of non-rigid structures, the ambiguity of camera position and 3D shape deformation is the main cause of the failure of posture recovery in real scenes. Although some effective solutions have been proposed, these methods are too restrictive for general 3D reconstruction and require too many cameras. So, to improve the problem of 3D shape and motion of non-rigid human body, Bastian, Hanno and Bodo [24] propose a potential solution, based on motion capture data, change real world data to factorize 2D observations in camera parameters, like basic poses and mixing coefficients, and achieve remarkable results.

Using a monocular camera to recover a 3D human pose can create an inherently ill-posed problem in the process of projecting a 3D image from 2D images. Aiming at improving the accuracy of 3D motion reconstruction, Du and Wong et al. [25] introduce the calculated height map into the previous algorithm for reconstructing the 3D motion under a single vision calibration camera.

Since reconstructing the configuration of 3D points from the projections in the image is an ill-posed problem, Ramakrishna et al. [26] use a motion-independent approach to solve this problem by using a large motion capture corpus as a proxy for visual memory. The 3D configuration of human body figure is restored from a 2D position with semantic meaning and anatomical landmarks in a single image.

Human body posture estimation is a key step in action recognition. During the 3D pose estimation, due to the lack of depth information, multiple 3D pose projections are likely to have the same 2D pose afterwards. And existing 2D pose detectors are often inaccurate, directly leading to errors in 3D pose estimation. Therefore, Wang et al. [27] learn the sparse and basic linear combination from 3D human bones to represent 3D poses and optimize them with alternating direction method (ADM), which greatly improve the accuracy of 3D pose estimation.

3.2. Machine learning and prediction method

In trajectory optimization, motion planning or grip estimation, it is difficult to accomplish them from a robot perspective. In order to enable a field such as a robot to truly understand a scene, Gupta et al. [28] first detect and segment objects in the scene, and then use a Convolutional Neural Network (CNN) to predict the pose of the object. Finally, the corresponding 3D model in the search library is aligned with the objects in the RGB-D scene.

In computer vision, the development of articulated pose estimation can be applied immediately to human tracking, motion recognition and video analysis. Many scholars have conducted detailed research on it. Most of their works in pose estimation are based on a graphical model. Chen et al. [29]
specify a graphical model of the human pose that can be used to detect components (or joints) and the spatial relationships between them (image-dependent pairwise relationships). The spatial relationship is then analyzed using a Deep Convolutional Neural Network (DCNN).

Accurate digitization of real-world objects is one of the core issues of visual computing. While the digitization scheme for rigid objects is widely available and can be considered a mature technique, the technique is no longer effective in the dynamics of the acquired shape being in motion and deformation. Because of many advantages of single-view 3D scanners, Li and Adams et al. [30] introduce a novel template-based dynamic registration algorithm to optimize the geometric details which are corrupted in a single view due to object motion.

3D human posture recovery is the most basic step in human motion recognition. The direct recovery of the pose by projection is usually ambiguous, so 3D recovery from a single view is slow. Lv F and Nevatia R [31] propose a method that does not explicitly infer 3D poses in every frame. In this approach, each action is modeled as a series of composite 2D human poses rendered from various viewpoints, and then a specific algorithm is used to form the optimal motion sequence.

3.3. Estimating body pose configuration from depth and silhouette

General marker-less MoCap studies typically focus on the use of one or more single view cameras, whereas M. Ye et al. [32] match a single depth map to a set of pre-captured motion samples to generate a body pose configuration which is succeeded to solve the problem emerged in the marker-less MoCap studies. In addition, they modified the point cloud smoothing technique to handle very noisy input depth maps.

The film industry or motion capture products for computer games are often marker-based. Although the accuracy of the marker-less method is comparably less to that of the marker-based method, the segmentation step is very restrictive to the capture environment because marker-less method relies on segmentation of people in the foreground. Grest et al. [33] do not use segmentation techniques and they track human motion by combining depth and contour information obtained from a single view.

3.4. Method of capturing joint movement

Although many commercial sensors have been developed that can detect 3D fingertip positions without the use of markers, these sensors are unable to restore the semantically significant skeleton type of the hand. In order to solve this problem, Sridhar et al.’s method [34] combines discriminant features based on local optimization, partial pose retrieval methods and pose estimation methods to capture various joint hand movements at an interactive rate.

4. Multi-view methods

4.1 Factorizing 2D observations to obtain convincing 3D reconstructions

Because multi-view systems can greatly reduce occlusion and blur when tracking joint structures such as human or hand, a faster and more powerful algorithm can be proposed based on this feature. Also, in multi-view conditions, the high dimensionality of the state space (many degrees of freedom) is also challenging. Kehl et al. [35] develop a novel approach to reconstructing the human body from contours in multiple video images. Whole body posture tracking is then performed by using random sampling.

Although the recovery of 3D human posture has made considerable progress in the past few years, the recovery of human posture still does not solve the problems of unconstrained movement in dynamic and messy environments. To solve these problems, Hofmann and Gavrila [36] use a framework to perform unconstrained 3D human upper body pose estimation from multiple camera views in a complex environment. The main contributions of this method are single frame pose recovery, time integration and model texture adaptation.
4.2 Machine learning and prediction method

When performing marker-less human motion tracking indoors or outdoors, it is likely to confront with the problems that the costs are expensive, the number of cameras is too large, and the accuracy of the tracking model is reduced. Elhayek, Aguiar et al. [2] combine the joint detection method based on discriminant image with the model-based motion tracking algorithm to solve these problems, and use Convolution Network (ConvNet) to accurately capture the joint skeleton motion of the subject in indoor and outdoor general scenes.

4.3 Method of capturing joint movement

Most methods for implementing mark-free motion capture assume that the cameras are synchronized, pre-calibrated, and static. To achieve these three points, specialized and expensive hardware will be required. Hasler, and Rosenhahn et al. [37] implement a method for marker-less motion capture for articulated objects that requires recording with multiple asynchronous handheld mobile cameras.

5. Discussion

As Ronald Poppe [38] puts it, human motion analysis is still an arduous problem due to the huge changes in camera movement and appearance, camera angle and environment settings. The key to analyzing human motion successfully is to effectively use the known information, like object appearance and movement. Since the end of the 19th century, scientists have analyzed human movements and have done a lot of researches. In the past two decades, due to the rapid development of computer hardware and software, human body motion analysis has become more and more precise and perfect. The human motion models originally described in 2D have now evolved into highly articulated 3D models. Mature commercial marker-based human motion capture models are also slowly being replaced by marker-less human motion capture models. In recent years, machine learning plays an increasingly important role in human motion analysis and makes great progresses, which will continue for decades.

For each of the methods and applications described in this survey, the problems encountered in the experiment were solved relatively perfectly. However, there are still some shortcomings that need further researches and developments in the future. For example, M. Ye et al. [32] are looking forward to spending more time in accelerating the calculations, and they also plan to invest in non-optical MoCap systems for more quantitative comparisons with Kinect [39]. M. Ye et al. hope that the single camera MoCap solution will play its important role in more areas. The method of Elhayek, Aguiar et al. [2] still has some shortcomings. Using their methods, motion tracking of a single camera view is not practicable. Moreover, the camera's frame rate needs to be sufficient to cope with the speed of the recorded motion. In this case, the cost will be greatly improved. So as a future job, Elhayek, Aguiar et al. [2] want to investigate the use of unsynchronized or moving cameras in their framework.

Therefore, some problems of the human motion model, their role and how they are extended to a wider range of areas contribute to human life require further researches.

References

[1] Moeslund T B and Granum E 2001 Computer Vision and Image Understanding. A survey of computer vision-based human motion capture, 81(3), 231-268.
[2] Elhayek A, Aguiar E D, Jain A, Tompson J, Pishchulin L and Andriluka M, et al. 2015 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. IEEE Computer Society.
[3] Ceseracciu E, Sawacha Z and Cobelli C 2014 PloS one. Comparison of marker-less and marker-based motion capture technologies through simultaneous data collection during gait: proof of concept. 9(3), e87640
[4] Zhou H and Hu H 2008 Biomedical Signal Processing and Control. Human motion tracking for rehabilitation——A survey. 3(1), 1-18
[5] Shingade A and Ghotkar A 2014 *International Journal of Computer Graphics & Animation*. Animation of 3D human model using marker-less motion capture applied to sports. 4(1), 27-39.

[6] Liu Y, Gall J, Stoll C, Dai Q, Seidel H P and Theobalt C 2013 *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Marker-less motion capture of multiple characters using multi-view image segmentation. 35(11), 2720-2735.

[7] D J Sturman 1999 *Character Motion Systems*. A brief history of motion capture for computer character animation. SIGGRAPH 94: Course 9.

[8] Field M, Pan Z, Stirling D and Naghdy F 2011 *Industrial Robot: An International Journal*. Human motion capture sensors and analysis in robotics. 38(2), 163-171.

[9] Fleisig G S 1999 *J. Biomech*. Kinematic and kinetic comparison of baseball pitching among various levels of development. 32.

[10] Riley P O, Paolini G, Croce U D, Paylo K W and Kerrigan D C 2007 *Gait and Posture*, A kinematic and kinetic comparison of overground and treadmill walking in healthy subjects. 26(1), 0-24.

[11] Gavrila D M and Davis L 1996 *Conference on Computer Vision & Pattern Recognition*. IEEE Computer Society. 3D model-based tracking of humans in action[C].

[12] Moses Y and Ullman S 1992 In: *Sandini G. (eds) Computer Vision — ECCV’92. ECCV 1992. Lecture Notes in Computer Science*. Limitations of non model-based recognition schemes. vol 588. Springer, Berlin, Heidelberg

[13] Mian A, Bennamoun M and Owens R 2007 *IEEE Transactions on Pattern Analysis and Machine Intelligence*. An efficient multimodal 2D-3D hybrid approach to automatic face recognition. 29(11), 1927-1943.

[14] Shafer S A, Stentz A and Thorpe C E 1986 *Proc. IEEE Conf.on Robotics & Automation*. An architecture for sensor fusion in a mobile robot. 3, 2002-2011.

[15] Tsitsiklis J N 1988 *Math. Control Signal Systems*. 1: 167.

[16] Cheu R L, Lee D H and Xie C 2002 *Intelligent Transportation Systems*. An arterial speed estimation model fusing data from stationary and mobile sensors. IEEE.

[17] Gray D, Brennan S and Tao H 2007 Evaluating Appearance Models for Recognition, Reacquisition, and Tracking.

[18] Mykhaylo Andriluka, Stefan Roth and Bernt Schiele 2010 *Computer Vision & Pattern Recognition*, Monocular 3D pose estimation and tracking by detection. 23(3), 623-630.

[19] Aldoma A, Marton Z C, Tombari F, et al. 2012 *IEEE Robotics & Automation Magazine*. Tutorial: point cloud library: three-dimensional object recognition and 6 dof pose estimation. 19(3), 80-91.

[20] Cao Z, Simon T, Wei S E and Sheikh Y 2016 Realtime multi-person 2D pose estimation using part affinity fields.

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

[22] Li R, Yang M H, Sclaroff S and Tian T P 2006 *European Conference on Computer Vision*. Monocular Tracking of 3D Human Motion with a Coordinated Mixture of Factor Analyzers. Springer, Berlin, Heidelberg.

[23] Field M, Pan Z, Stirling D and Naghdy F 2011 *Industrial Robot: An International Journal*. Human motion capture sensors and analysis in robotics. 38(2), 163-171.

[24] Wandt B, Ackermann H and Rosenhahn B 2015 *IEEE Conference on Computer Vision and Pattern Recognition*. 3D Human Motion Capture from Monocular Image Sequences. IEEE.

[25] Du Y, Wong Y, Liu Y, Han F and Geng W 2016 *European Conference on Computer Vision*. Marker-Less 3D Human Motion Capture with Monocular Image Sequence and Height-Maps. Springer International Publishing.

[26] Ramakrishna V, Kanade T and Sheikh Y 2012 *Proceedings of the 12th European conference on Computer Vision*. Reconstructing 3D Human Pose from 2D Image Landmarks. Springer, Berlin, Heidelberg. Volume Part IV.
[27] Wang C, Wang Y, Lin Z, Yuille A L and Gao W 2014 Robust estimation of 3d human poses from a single image.
[28] Gupta S, Arbelaez P, Girshick R and Malik J 2015 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Aligning 3D models to RGB-D images of cluttered scenes. IEEE Computer Society.
[29] Chen X and Yuille A L 2014 In NIPS. Articulated pose estimation by a graphical model with image dependent pairwise relations.
[30] Li H, Adams B, Guibas L J and Pauly M 2009 ACM Transactions on Graphics, [acm press acm siggraph asia 2009 papers - yokohama, japan (2009.12.16-2009.12.19)] acm siggraph asia 2009 papers on - siggraph asia "09 - robust single-view geometry and motion reconstruction. 1.
[31] Lv F and Nevatia R 2007 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA. Single View Human Action Recognition using Key Pose Matching and Viterbi Path Searching. IEEE.
[32] Ye M, Wang X, Yang R, Ren L and Pollefeys M 2011 IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011. Accurate 3D Pose Estimation From a Single Depth Image. IEEE.
[33] Grest D, Krüger, Volker and Koch R 2007 Lecture Notes in Computer Science, Single view motion tracking by depth and silhouette information. 4522(2), 719-729.
[34] Sridhar S, Oulasvirta A and Theobalt C 2013 Proceedings of the 2013 IEEE International Conference on Computer Vision. Interactive Markerless Articulated Hand Motion Tracking Using RGB and Depth Data. IEEE.
[35] Kehl R, Bray M and Gool L J V 2005 IEEE Computer Society Conference on Computer Vision & Pattern Recognition. Full Body Tracking from Multiple Views Using Stochastic Sampling. IEEE.
[36] Hofmann M and Gavrila D M 2012 International Journal of Computer Vision, Multi-view 3d human pose estimation in complex environment. 96(1), 103-124.
[37] Hasler N, Rosenhahn B, Thorsten Thormählen, Michael W, Gall J and Seidel H, et al. 2009 IEEE Conference on Computer Vision & Pattern Recognition. Markerless motion capture with unsynchronized moving cameras. IEEE.
[38] Poppe R 2007 Computer Vision and Image Understanding. Vision-based human motion analysis: an overview. 108(1-2), 4-18.
[39] Microsoft 2010 Kinect camera. http://www.xbox.com/en-US/kinect/default.htm, 1, 2