Tracking of hands holding objects: Implementations and limitations of techniques and methods

Claudiu Andrei - 180015483
ec211193@qmul.ac.uk

***Abstract*—Real time information extraction from data in video format poses a significant challenge in the field of computer vision. Tracking of hands manipulating objects in a three-dimensional environment presents a plethora of approaches and manipulation of difficult data and structures, however the potential applications in segments such as wearable computers, augmented reality and in some cases sports performance analysis evince the importance and the prospects of this practice, incentivizing further research. This paper proposes an insight on some extensive literature review carried out on existing proposed methodologies and techniques offered by various computer vision enthusiast. Given the number of different types of possible approaches when tracking the manipulation of objects in real time, this paper analyses the implications of some popular pipelines.**

I. Introduction

Hands are a very difficult human feature to precisely track; extrapolating their behaviour in real time and tracking their movement already proves challenging because of the human anatomy: with 27 bones, 27 joints, 34 muscles and more than 100 between ligaments and tendons it quickly becomes evident that defining a system able to accurately model all these variables is not a straightforward task. While it is not uncommon for items of daily use such as smartphones or cars to track the movement and behaviour of hands by recognizing certain gestures, many of the contributions and existing research do not focus on the interaction between articulated three-dimensional models and objects. In the deep learning area, many approaches have been explored vastly, with some achieving very successful results such as Depth-based approaches [1][2][3][4][5] and RGB based approaches [6][7][8]. As previously mentioned, the success these methods prove often does not transfer to the ability of tracking hands in manipulation tasks: this is usually the case because of the occlusion that the object cause, blocking the view of the hand and impeding accurate tracking.

II. Problem Definition

As mentioned previously the key difficulties that are encountered when approaching this task mainly emerge because of the occlusions created by objects when being manipulated. Many proposed solutions benefit from the use of RGBD formats, getting an extra advantage from having information about the depth of the data presented by cameras which operate using a depth sensor [9]. Other approaches of 3D hand tracking problem can be formulated by outlining an optimisation where the solution is a model configuration aimed at the maximisation of the colour consistency in 2 different camera views, also known as RGB-stereo methods [10].

III. Key Works

Following, is a discussion of existing key works in the area. Approaches with robust estimation performance exist and have been explored thoroughly, in this document however, a handful have been shortlisted based on their relevance, available supplementary data and, in order to give insight on diverse methods, their type of approach. The different methodologies revolve around different observation models. [11] Uses a system with range data of resolution 640 by 480 depth points captured at 25Hz merged with conventional two-dimensional RGB colour space images. This approach is defined as depth-based: it uses colour information exploitation in a pre-processing step to locate the hand and detect occlusions caused by the object being held via skin colour segmentation. Here the different segments of the hand have their own models and local tracker, which draws a number of samples from a proposal function at every computation step. This method exploits anatomical constraints to enforce pre-defined hand structures by defining a Markov random field where each one of the local hand trackers corresponds to a rigid hand segment. [12] Proposes a different methodology, in fact here the observation of the scene is carried out using a pair of RGB cameras, therefore the depth attribute is not present here, however, thanks to the dual camera setup, it is possible to estimate the position of the hand by defining a colour consistency quantification system built by employing Particle Swarm Optimization (PSO)[13]. The anatomically consistent and realistic hand models utilized here are provided by lib-hand[14], a skinned model consisting of 22 bones. This solution casts the problem of hand tracking as an optimization task where maximization of colour consistency is pursued.[15] An interesting and unique approach is taken here: in fact this method takes an input a monocular RGB image, therefore having access to the least amount of data amongst the methods observed so far. The solution proposed here revolves around a robust non-parametric methos for 3D hand reconstruction, which operates in real time and accounts for time continuity constraints. The method performs for an image, a nearest neighbor search in a database of 100.000 hand poses and is able to run at around 10 FPS (frames per second). For each frame or specific time instant, a 28 dimensional vector is defined for the model of joint angles while a different 512D histogram of oriented gradients (HOG) represents the observation constraints.

IV. Evaluation Criteria

In order to evaluate the efficiency, reliability and real-world practicality of these methods it is important to understand the context in which they are to be utilized. Depending on the available hardware and systems different methods must be chosen. In an industrial environment, where a company might want to implement hand object tracking, a moderately accurate approach might be chosen, offering a good balance between precision and requirements. In applications where the tracking is to be used by general population, the factor that must be taken into account is the availability of data provided: not everyone might have a smartphone with a depth sensor providing and RGBD image, therefore a method which relies solely on RGB data would prove mor practical in this instance. Professional use of these tools is also an important application: in the sector of physiotherapy and rehabilitation a very accurate model is probably required, and more advanced tools are available for carrying out the tracking in the most accurate way. These are some of the variables that need to be considered when evaluating these methodologies, therefore the success of each

can differ based on the application. As a rule of thumb, if the requirements are met or surpassed for a certain application, the methods are viable for the designated task.

V. Discussion

This is the core section of the report. Provide an analysis of the literature/results.[11] Offers a configuration of local parts coupled by soft constraints, rather than the complete hand model, and to explicitly model occlusion by both hand parts and objects. Valid hand configurations are enforced by means of a MRF model connecting the segments. The method proves relatively effective in certain controlled scenarios however its expensive computational time of 6.2 seconds per frame is enough to eliminate it from almost every real time application. In certain scenarios it would prove effective for analysis of existing data sets, moreover, as these must be constructed in a controlled environment, the process would prove meticulous and hard to repeat for real world applications. This type of pipeline would prove useful in analysis of sports performance such as datasets portraying athletes in baseball, bowling and possibly darts. It is discussed previously that various approaches prove to be valid in different scenarios; [12] for example, appears to be a better solution than the one mentioned above. Here near real time performance is achieved using a dedicated graphics processor: 15-20 FPS. The model accuracy is comparable to the performance achieved by methods which take advantage of contemporary RGBD cameras providing information about depth in the scene, furthermore, this method is able to recognize three scenarios consisting of: one hand, object and hand, and two hands sequences. This appears to be the most balanced pipeline as it offers great performance, low computational cost, and an easier method of capturing data. Given its real time and precise nature, this approach might have more practical applications than [11], such as motion tracking while interacting with objects and can be further developed to account for problems such as 3D tracking of rigid objects and human skeleton tracking. [15] Is a project that appears to offer real life applications for the general public by being possibly implemented in mobile apps or video game consoles. Its main advantage is the use of a simple setup, meaning it can be employed by everyone, therefore allowing for further development and research. Its accessibility however has a constraint when it comes to implementation: its database needs to be carried/imported in every system where it needs to be used. Moreover, the amount of recognizable poses is linked to the number of unique actions recorded in the database. [15] Is probably the most likely to be improved as it could easily be employed by large companies such as Google, which would be able to store several million unique poses on their cloud platform and have a mobile application that is able to offer a state-of-the-art hand and object tracking experience by implementing a pipeline similar to that of their voice assistant.

VI. Conclusion

As seen all of these approaches would appear to be fairly efficient and practical at first glance, however when analyzing them in depth and choosing one for a certain application it quickly becomes evident that many constraints are the limiting factors. This leaves room from improvement and further research might be carried out to optimize current pipelines or identify new techniques: as seen most methods require the use of complex set ups and are computationally expensive. A fast and reliable method available to the public would open new doors for mobile applications development and would allow medical advancements with uses in pen and hand tracking for people affected by disability.

References

[1] S. Yuan, G. Garcia-Hernando, B. Stenger, G. Moon, J. Chang, K. Lee, P. Molchanov, J. Kautz, S. Honari, L. Ge, et al. Depth-based 3d hand pose estimation: From current achievements to future goals. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2636–2645, 2018

[2] L. Ge, Y. Cai, J. Weng, and J. Yuan. Hand pointnet: 3d hand pose estimation using point sets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8417–8426, 2018.

[3] L. Ge, H. Liang, J. Yuan, and D. Thalmann. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3593–3601, 2016.

[4] L. Ge, H. Liang, J. Yuan, and D. Thalmann. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1991–2000, 2017.

[5] M. Oberweger, P. Wohlhart, and V. Lepetit. Hands deep in deep learning for hand pose estimation. arXiv:1502.06807, 2015.

[6] A. Boukhayma, R. Bem, and P. Torr. 3d hand shape and pose from images in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 10843–10852, 2019.

[7] P. Panteleris, I. Oikonomidis, and A. Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 436–445. IEEE, 2018.

[8] L. Yang, S. Li, D. Lee, and A. Yao. Aligning latent spaces for 3d hand pose estimation. In Proceedings of the International Conference on Computer Vision, 2019

[9] P. Panteleris, N. Kyriazis, and A. A. Argyros. 3d tracking of human hands in interaction with unknown objects. In BMVC, 2015.

[10] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In ICCV, 2011.

[11] H. Hamer, K. Schindler, E. Koller-Meier and L. Van Gool, "Tracking a hand manipulating an object," 2009 IEEE 12th International Conference on Computer Vision, 2009, pp. 1475-1482, doi: 10.1109/ICCV.2009.5459282.

[12] Panteleris, P., & Argyros, A.A. (2017). Back to RGB: 3D Tracking of Hands and Hand-Object Interactions Based on Short-Baseline Stereo. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 575

[13] M. Clerc and J. Kennedy. The particle swarm - explosion, stability, and convergence in a multidimensional complex space. Transactions on Evolutionary Computation, 2002.

[14] M. Saric. Libhand: A library for hand articulation. http://www.libhand.org/, 2011. Version 0.9.

[15] J. Romero, H. Kjellstrom, and D. Kragic. Hands in action: ¨ real-time 3d reconstruction of hands in interaction with objects. In ICRA, 2010.