A Generative Approach to Tracking Hands and Their Interaction with Objects

Nikolaos Kyriazis, Iason Oikonomidis, Paschalis Panteleris, Damien Michel, Ammar Qammaz, Alexandros Makris, Konstantinos Tzevanidis, Petros Douvantzis, Konstantinos Roditakis and Antonis Argyros

Abstract Markerless 3D tracking of hands in action or in interaction with objects provides rich information that can be used to interpret a number of human activities. In this paper, we review a number of relevant methods we have proposed. All of them focus on hands, objects and their interaction and follow a generative approach. The major strength of such an approach is the straightforward fashion in which arbitrarily complex priors can be easily incorporated towards solving the tracking problem and their capability to generalize to greater and/or different domains. The proposed generative approach is implemented in a single, unified computational framework.

Keywords 3D hand tracking · 3D tracking of hand-object interactions · Model-based 3D tracking

1 Introduction

This work discusses the problem of observing and understanding human (inter)action through markerless observations, i.e. in a non noninvasive manner. Human action is considered with focus on human hands. Understanding human hands in action is a theoretically and practically interesting problem. Humans are readily capable of understanding hand actions of their own and of other humans. Interestingly, the idea supported in the literature is that to achieve such understanding, humans employ mental simulation [6, 7], which could tentatively be corresponded to the generative approach in computer vision. On the practical side, achieving the understanding

N. Kyriazis · I. Oikonomidis · P. Panteleris · D. Michel · A. Qammaz
A. Makris · K. Tzevanidis · P. Douvantzis · K. Roditakis · A. Argyros (✉)
Institute of Computer Science, Foundation for Research and Technology, Heraklion, Greece
e-mail: argyros@ics.forth.gr

N. Kyriazis · I. Oikonomidis · K. Tzevanidis · P. Douvantzis · K. Roditakis · A. Argyros
Computer Science Department, University of Crete, Heraklion, Greece

A. Gruca et al. (eds.), *Man–Machine Interactions 4*, Advances in Intelligent Systems and Computing 391, DOI 10.1007/978-3-319-23437-3_2

of human hand action sets the foundation upon which a variety of socially helpful applications can be established, in the fields of safety, medicine, education, industry, entertainment, etc. Substituting occasionally flawed visual scrutiny by humans for systematically robust mechanised processing of images can have a significant positive impact on the success of the underlying tasks.

As demonstrated in our work, rich understanding of hand action and interaction with objects can be successfully built upon robust and detailed 3D tracking of hands and objects, from images captured by noninvasive camera setups. However, the 3D tracking task is not an easy one. The structural complexity of the human hand corresponds to multiple Degrees of Freedom (DoFs) which yield a tracking problem with a difficult to explore high-dimensional search space. Highly frequent and temporally prolonged self-occlusions lead to insufficient observations for full and proper estimation of hand articulation. The uniformity of finger appearance introduces ambiguities in observation, where e.g. one finger can be mistaken for another. It is common that human hands occupy little area in captured images and therefore correspond to observations of low spatial resolution. On top of that, hands may move quite fast, resulting in a relatively low temporal resolution, too, even if observed by cameras operating in 30 fps. Adding more hands and/or objects in 3D tracking only aggravates the aforementioned problems and also increases computational complexity.

2 The Proposed Generative Approach

We present our generative approach in problems pertaining to 3D tracking of hands in isolation Sect. 2.1 and hands in interaction with objects Sect. 2.2. A unified approach [11] is employed in all of the presented work.

The presented work is founded on the premise that there exists a simulation process which can synthesize data similar to acquired observations. The configuration of the simulation that best matches some observations is the "explanation" or understanding of the observations. Simulations involve synthesizing appearance, through 3D rendering, and even behaviour, if physics-based simulation is employed.

Assuming temporal continuity, tracking objects in 3D amounts to searching for the 3D configuration of the objects whose simulation best matches actual observations. Searching is performed in the vicinity of the tracking result for the previous frame. For each tracking frame the simulation error $\mathbf{E}$ is minimized:

$$\mathbf{E}(x, o, h) = \|\mathbf{M}(x, h) - \mathbf{P}(o)\| + \lambda \mathbf{L}(x, h), \tag{1}$$

where x is the state of all tracked objects, o are the observations of the current frame, h is the tracking history, $\mathbf{M}$ is the simulation function, $\mathbf{P}$ is the pre-processing function which maps observation to the same feature space as $\mathbf{M}$ and $\mathbf{L}$ is a prior on the 3D configuration of objects. The first term is a notational abstraction of a data term, i.e. a quantification of the discrepancy between a given hypothesis and actual

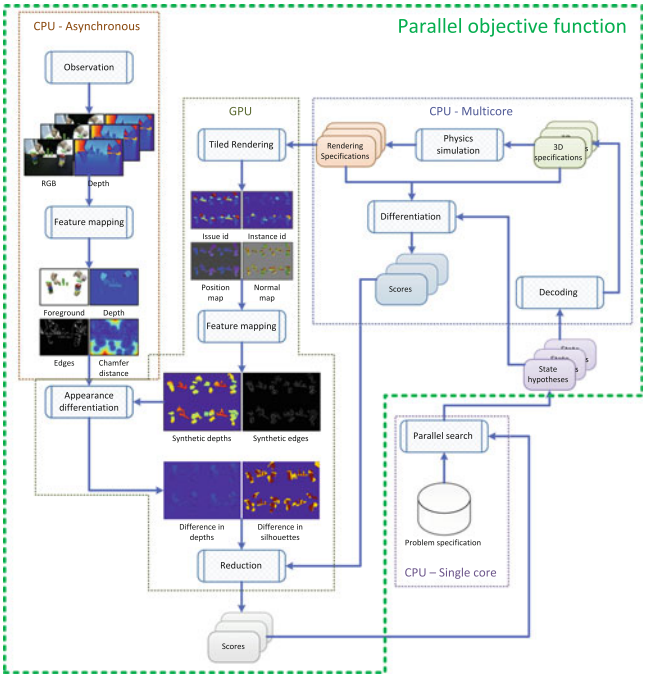

Fig. 1 The outline of the implementation for the generative 3D tracking framework. For more details the reader is referred to [11]

measurements. The prior term and data term are balanced with a weight λ that is empirically defined.

The tracking solution s for the current frame amounts to minimizing (1):

$$s \triangleq \arg\min_{x} \mathbf{E}(x, o, h). \tag{2}$$

During search for the optimal x, (1) is invoked several times. To favor speed, parallel search techniques are incorporated and the implementation of (1) exploits GPU acceleration. An outline of the implementation of the generic approach is depicted in Fig. 1.

2.1 Tracking Hands in Isolation

There is significant interest in pursuing fast and robust 3D tracking of a single hand. However, as already discussed, tracking a single hand in 3D is already a formidable challenge. We hereby present a series of successful 3D single hand tracking efforts.

Relevant Work The problem of 3D hand tracking has received significant attention. There exist several approaches to solving the problem, with every one including a top-down component and a bottom-up component. According to the contribution weight of each component, approaches can be characterized as top-down (dominating top-down component), bottom-up (dominating bottom-up component) and hybrid (equally contributing top-down and bottom-up components). Top-down methods work by establishing a model of the hand and fitting it to moderately preprocessed observations, through search [3, 14, 29, 31, 35, 38, 42, 43, 48]. The methods presented in this paper call within this category too. Bottom-up methods work by learning, via machine learning tools, the mapping from the image space to the configuration space, from large training corpora [1, 4, 8, 9, 16, 34, 36, 44, 46]. Hybrid methods work by fusing top-down models with strong bottom-up features, which are a product of learning [30, 40, 41, 49].

Contribution We have investigated the 3D tracking problem for a hand in isolation and under different cases of camera setups, representations and optimization tools [5, 15, 20–22, 24].

With respect to camera setups we have explored two options, namely a set of calibrated RGB cameras and a single RGBD camera (MS Kinect, ASUS Xtion). What is differentiated between the two classes is the feature space used while computing (1). For the case of multiple RGB cameras we have employed per camera 2D features (silhouettes, edges) which are fused in 3D information as a result of incorporation in an objective function (see 1) defined over 3D configuration [24]. We have also employed 3D features (visual hulls) which, conversely, where computed as a fusion of multiple 2D cues prior to their incorporation in the computations of the objective function [21]. The latter [21] presents favorable traits for the small baseline stereo (2 cameras) case and is otherwise similar to the former [24]. For the RGBD case we have employed both 2D (silhouettes) and 3D (depth) features in (1) [5, 15, 20, 22]. The benefit of employing a single RGBD camera instead of multiple RGB cameras is the significant decrease of data which require processing while maintaining the required 3D information. This allows for 3D tracking of a single hand which can currently be performed at a rate of 30 fps.

The default representation of a hand in all presented work has been a kinematics tree (skeleton) of 27 parameters, redundantly encoding 26 DoFs. However, it is noticed that hand motion is often modulated by an underlying task. We investigated whether by conditioning on tasks, hand motion could be described with a reduced amount of parameters. In [5] we came up with per-task sub-spaces of hand motion, which we successfully employed in 3D hand tracking.

Last but not least, we have employed different optimization schemes for computing (2). In most of the hand tracking work [5, 20, 21, 24] (2) was computed using a variant (see [20] for details) of the Particle Swarm Optimization algorithm [28]. Introducing a custom evolutionary technique for search [22] and replacing PSO with it dramatically decreased the amount of required invocations to (1) for computing (2), with respect to [20], while preserving the accuracy level. By substituting the find-best optimization schemes of [20, 22] with a probability density propagation scheme,

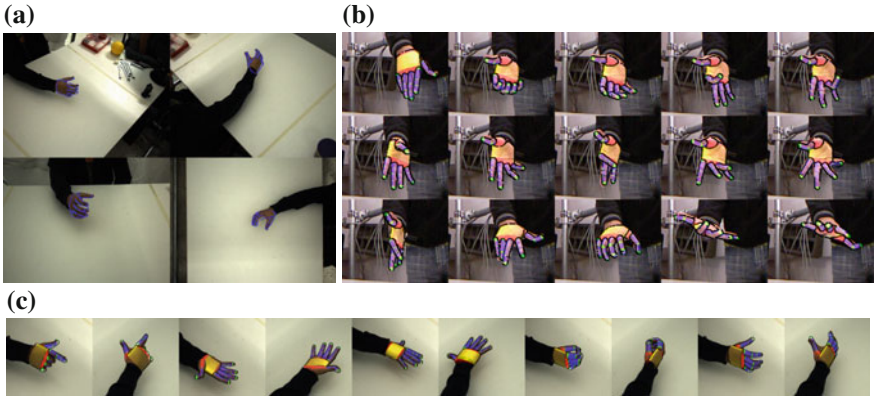

Fig. 2 Exemplar results of 3D tracking of a single hand from **a** multiple 2D cues [24], **b** a single RGBD camera [20] and **c** convex hulls computed from multiple views [20]

namely Hierarchical Particle Filtering [15], not only 3D hand tracking became even faster (90 fps) but it also became more robust, due to the persistence of several hypotheses across frames, instead of a single one as in [5, 20, 22, 24]. Indicative results of the 3D hand tracking methods are provided in Fig. 2. For more details the reader is referred to the corresponding papers.

2.2 Tracking Interacting Hands

In a significant subset of the scenarios which involve observation of a human, hands are not isolated. They are usually found interacting with other hands or with objects in the surrounding of the subject. The amount of objects manipulated over time can vary from one (explore or use a single object) to many (preparing multi-ingredient food). Tracking such scenes in 3D inherits the list of single hand tracking problems, in aggravated forms, accentuated by the cardinality of the scenes.

Relevant Work There exist some approaches in the literature to handle 3D tracking of hand-object interactions, that can be corresponded to the categorization of top-down [47], bottom-up [33] and hybrid [2] methods. However, they all regard a single object. On the other hand, there exist approaches that tackle the problem of tracking multiple objects in 3D, but do not include a hand [10, 37, 45].

Contribution We have performed work which spans across several choices over setups, representations, etc. Even more importantly, the corresponding methods also regard different scene scales. We will present the corresponding methods focusing on the axis of scene scale and denoting important points on the way.

The multi-camera 3D hand tracking method of [24] has been extended in [19] to also include an object of known parametric shape. The extension amounted to introducing the DoFs of the object and its 3D shape in the computations of (1) and (2). Thus, both the hand and the object where tracked, by requiring a computational budget which was a bit larger than that used in [24]. In [23] we showed that this extension can successfully scale up to the case of the object being as complex as another hand, increasing the DoFs to 54 (27 for each hand) and also increasing the computational budget. In both cases, the state of the hand and the other hand or object were considered jointly, as required in order to model constraints such as, for example, that two physical objects cannot occupy the same physical space. PSO was used in both cases and succeeded, given enough computational budget.

However, generalizing to larger scenes is problematic, as the total DoFs dramatically increase. Because of lack of gradient or any other type of focus, PSO cannot effectively explore arbitrarily large search spaces with limited resources. To introduce a sense of focus, the joint problem of tracking many objects and/or hands in 3D was carefully decomposed to several semi-independent tracking problems, one for each entity [13]. Instead of a single tracker for the entire scene, an Ensemble of Collaborative Trackers is incorporated. Each of these trackers treats the last known state of all others as static and thus becomes independent of them, while at the same time it incorporates their state in its computations. This allows to simultaneously decompose the problem but also preserve reasoning over the entire state in each tracker. The decomposition, along with the parallel implementation (CPU and GPU) yield sub-linear increase in computational time as the number of objects increases and significantly outperforms schemes such as the joint optimization of [19, 23] and truly independent trackers (e.g. multiple instances of [20]).

The ECT method in [13] was also endowed with a hand-object manipulation prior which allowed for even computing forces from vision alone [27]. When a hand and an object are tracked by an ECT, erroneous measurements or optimization hiccups can lead to physically implausible manipulation estimations. By tracking the acceleration of the manipulated object and consulting a per finger force prior (learned through training) the hand-object estimation can be rectified so as to become physically plausible. At the same time, the actual forces exerted by the subject are also computed.

The notion of physical plausibility has also been exploited in [12] to enable scalable tracking of a scene comprising a hand and several objects. The observation that objects move only due to the motion of the hand leads to the formulation of a hand-objects tracking problem whose DoFs are the same as the hand's alone. Otherwise stated, the amount of passive objects makes no difference in the size of the search space, since in order to track any subset of moving objects it suffices to only search for a hand motion whose physical consequences would have the objects move. Some results of the 3D hand-object(s) tracking methods can be viewed at Fig. 3. For more details the reader is referred to the corresponding papers.

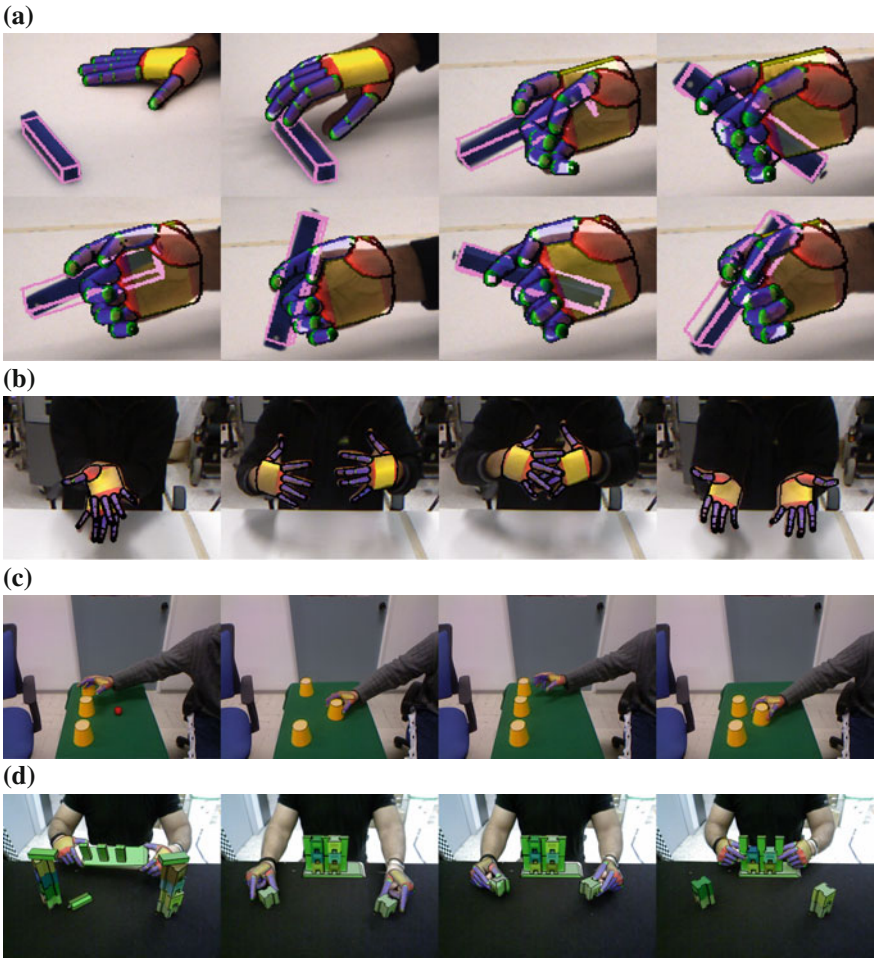

Fig. 3 Exemplar results of 3D tracking of hands and objects in interaction: **a** hand-object interaction observed from a multi-camera system [18], **b** two strongly interacting hands observed from an RGBD camera [23], **c** a hand interacting with a few objects observed from an RGBD camera [12] and **d** two hands interacting with many objects observed from an RGBD camera [13]

3 Discussion

We presented a series of methods to solving the problem of tracking hands in action or in interaction with objects, in 3D. All the works followed the generative approach, as implemented in a single unified and modular architecture [11]. The generative approach has successfully led to tackling the 3D tracking problems and establishing state-of-the-art solutions. At the same time, the generative approach, as opposed to discriminative or hybrid approaches, has straightforward means to facilitate

generalization to larger and/or different domains. As an example, the same approach used in the presented work has also been successfully employed in problems of tracking hands with markers [32], tracking head motion [25] and tracking the full body [17]. The results of the presented 3D hand-object tracking methods have also been employed to fuel even higher-level inference. Specifically, reconstructed and detailed 3D trajectories of hands manipulating objects have been used to establish a high-level action grammar for everyday tasks [26] and also to enable the inference of the manipulation intent a subject has while approaching an object, even before manipulation is actually observed [39].

Acknowledgments This work was partially supported by the by the EU ISTFP7-IP-288533 project RoboHow.Cog and by the EU FP7-ICT-2011-9-601165 project WEARHAP.

References

1. Athitsos, V., Sclaroff, S.: Estimating 3d hand pose from a cluttered image. In: CVPR 2003. vol. 2, pp. 432–439. Madison, USA (2003)
2. Ballan, L., Taneja, A., Gall, J., Van Gool, L., Pollefeys, M.: Motion Capture of Hands in Action Using Discriminative Salient Points. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 640–653. Springer, Heidelberg (2012)
3. Bray, M., Koller-Meier, E., Van Gool, L.: Smart particle filtering for high-dimensional tracking. Comput. Vis. Image Underst. **106**(1), 116–129 (2007)
4. de Campos, T., Murray, D.: Regression-based hand pose estimation from multiple cameras. In: CVPR 2006. vol. 1, pp. 782–789. New York, USA (2006)
5. Douvantzis, P., Oikonomidis, I., Kyriazis, N., Argyros, A.: Dimensionality reduction for efficient single frame hand pose estimation. In: Chen, M., Leibe, B., Neumann, B. (eds.) Computer Vision Systems. LNCS, vol. 7963, pp. 143–152. Springer, Heidelberg (2013)
6. Gallese, V., Goldman, A.: Mirror neurons and the simulation theory of mind-reading. Trends Cogn. Sci. **2**(12), 493–501 (1998)
7. Grezes, J., Decety, J.: Functional anatomy of execution, mental simulation, observation, and verb generation of actions: a meta-analysis. Hum. Brain Mapp. **12**(1), 1–19 (2001)
8. Keskin, C., Kirac, F., Kara, Y., Akarun, L.: Real time hand pose estimation using depth sensors. In: ICCV 2011. pp. 1228–1234. Barcelona, Spain (2011)
9. Keskin, C., Kıraç, F., Kara, Y., Akarun, L.: Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 852–863. Springer, Heidelberg (2012)
10. Kim, K., Lepetit, V., Woo, W.: Keyframe-based modeling and tracking of multiple 3D objects. In: ISMAR 2010. pp. 193–198. Seoul, Japan (2010)
11. Kyriazis, N.: A computational framework for observing and understanding the interaction of humans with objects of their environment. Ph.D. thesis, University of Crete (2014)
12. Kyriazis, N., Argyros, A.: Physically plausible 3D scene tracking: The single actor hypothesis. In: CVPR 2013. pp. 9–16. Portland, Oregon, USA (2013)
13. Kyriazis, N., Argyros, A.: Scalable 3D tracking of multiple interacting objects. In: CVPR 2014. pp. 3430–3437. Columbus, Ohio, USA (2014)
14. MacCormick, J., Isard, M.: Partitioned sampling, articulated objects, and interface-quality hand tracking. In: In: Proceedings of the 6th EuropeanConference on Computer Vision-Part II, ECCV 2000. LNCS, pp. 3–19. Springer, Heidelberg (2000)

15. Makris, A., Kyriazis, N., Argyros, A.: Hierarchical particle filtering for 3d hand tracking. In: CVPR 2015. pp. 8–17. Boston, USA (2015)
16. Malassiotis, S., Strintzis, M.G.: Real-time hand posture recognition using range data. Image Vis. Comput. **26**(7), 1027–1037 (2008)
17. Michel, D., Panagiotakis, C., Argyros, A.A.: Tracking the articulated motion of the human body with two RGBD cameras. Mach. Vis. Appl. **26**(1), 41–54 (2013)
18. Oikonomidis, I., Kyriazis, N., Argyros, A.: Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In: ICCV 2011. pp. 2088–2095. Barcelona, Spain (2011)
19. Oikonomidis, I., Kyriazis, N., Argyros, A., et al.: Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In: ICCV 2011. pp. 2088–2095. Barcelona, Spain (2011)
20. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Efficient model-based 3D tracking of hand articulations using kinect. In: BMVC 2011. p. 3. Dundee, UK (2011)
21. Oikonomidis, I., Kyriazis, N., Tzevanidis, K., Argyros, A., et al.: Tracking hand articulations: relying on 3D visual hulls versus relying on multiple 2D cues. In: ISUVR 2013. pp. 7–10. Daejeon, South Korea (2013)
22. Oikonomidis, I., Lourakis, M., Argyros, A., et al.: Evolutionary quasi-random search for hand articulations tracking. In: CVPR 2014. pp. 3422–3429. Columbus, Ohio, USA (2014)
23. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Tracking the articulated motion of two strongly interacting hands. In: CVPR 2012. pp. 1862–1869. Providence, Rhode Island (2012)
24. Oikonomidis, I., Kyriazis, N., Argyros, A.: Markerless and efficient 26-DOF hand pose recovery. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) Computer Vision—ACCV 2010. LNCS, vol. 6494, pp. 744–757. Springer, Heidelberg (2011)
25. Padeleris, P., Zabulis, X., Argyros, A., et al.: Head pose estimation on depth data based on particle swarm optimization. In: CVPRW 2012. pp. 42–49. Providence, USA (2012)
26. Patel, M., Ek, C.H., Kyriazis, N., Argyros, A., Miro, J.V., Kragic, D.: Language for learning complex human-object interactions. In: ICRA 2013. pp. 4997–5002. Karlsruhe, Germany (2013)
27. Pham, T.H., Kheddar, A., Qammaz, A., Argyros, A.A.: Towards force sensing from vision: observing hand-object interactions to infer manipulation forces. In: CVPR 2015. pp. 1893–1902. Boston, USA (2015)
28. Poli, R., Kennedy, J., Blackwell, T.: Particle swarm optimization. Swarm Intell. **1**(1), 33–57 (2007)
29. Qian, C., Sun, X., Wei, Y., Tang, X., Sun, J.: Realtime and robust hand tracking from depth. In: CVPR 2014. pp. 1106–1113. Columbus, USA (2014)
30. Qian, C., Sun, X., Wei, Y., Tang, X., Sun, J.: Realtime and robust hand tracking from depth. In: CVPR 2014. pp. 1106–1113. Ohio, USA (2014)
31. Rehg, J.M., Kanade, T.: Visual tracking of high DOF articulated structures: an application to human hand tracking. In: ECCV 1994. pp. 35–46. Springer, London, UK (1994)
32. Roditakis, K., Argyros, A.: Quantifying the effect of a colored glove in the 3D tracking of a human hand. In: Nalpantidis, L., Krüger, V., Eklundh, J.O., Gasteratos, A. (eds.) Computer Vision Systems. LNCS, vol. 9163, pp. 404–414. Springer, Heidelberg (2015)
33. Romero, J., Kjellström, H., Ek, C.H., Kragic, D.: Non-parametric hand pose estimation with object context. Image Vis. Comput. **31**(8), 555–564 (2013)
34. Romero, J., Kjellström, H., Kragic, D.: Monocular real-time 3D articulated hand pose estimation. In: IEEE-RAS 2009. pp. 87–92. Paris, France (2009)
35. Ros, G., del Rincon, J., Mateos, G.: Articulated particle filter for hand tracking. In: ICPR 2012. pp. 3581–3585. Stockholm, Sweden (2012)
36. Rosales, R., Athitsos, V., Sigal, L., Sclaroff, S.: 3D hand pose reconstruction using specialized mappings. In: ICCV 2001. vol. 1, pp. 378–385. Vancouver, Canada (2001)
37. Salzmann, M., Urtasun, R.: Physically-based motion models for 3D tracking: a convex formulation. In: ICCV 2011. pp. 2064–2071. Barcelona, Spain (2011)

38. Schmidt, T., Newcombe, R., Fox, D.: Dart: Dense articulated real-time tracking. In: RSS 2014. vol. 2. Berkeley, USA (2014)
39. Song, D., Kyriazis, N., Oikonomidis, I., Papazov, C., Argyros, A., Burschka, D., Kragic, D.: Predicting human intention in visual observations of hand/object interactions. In: ICRA 2013. pp. 1608–1615. Karlsruhe, Germany (2013)
40. Sridhar, S., Oulasvirta, A., Theobalt, C.: Interactive markerless articulated hand motion tracking using RGB and depth data. In: ICCV 2013. pp. 2456–2463. Sydney, Australia (2013)
41. Sridhar, S., Rhodin, H., Seidel, H.P., Oulasvirta, A., Theobalt, C.: Real-time hand tracking using a sum of anisotropic gaussians model. In: 3DV 2014. Tokyo, Japan (2014)
42. Stenger, B., Mendonça, P.R., Cipolla, R.: Model-based 3D tracking of an articulated hand. In: CVPR 2001. vol. 2, pp. 310–315. Kauai, HI, USA (2001)
43. Sudderth, E., Mandel, M., Freeman, W., Willsky, A.: Visual hand tracking using nonparametric belief propagation. In: CVPRW 2004. pp. 189–189. Washington, USA (2004)
44. Tompson, J., Stein, M., Lecun, Y., Perlin, K.: Real-time continuous pose recovery of human hands using convolutional networks. ACM Trans. Graph. **33**(5), 169–169 (2014)
45. Vo, B., Vo, B., Pham, N., Suter, D.: Joint detection and estimation of multiple objects from image observations. IEEE Trans. Signal Process. **58**(10), 5129–5141 (2010)
46. Wang, R., Paris, S., Popovic, J.: 6D hands: markerless hand-tracking for computer aided design. In: UIST 2011. pp. 549–558. UIST '11, ACM, New York, USA (2011)
47. Wang, Y., Min, J., Zhang, J., Liu, Y., Xu, F., Dai, Q., Chai, J.: Video-based hand manipulation capture through composite motion control. ACM Trans. Graph. (TOG) **32**(4), 43:1–43:14 (2013)
48. Wu, Y., Lin, J.Y., Huang, T.S.: Capturing natural hand articulation. In: ICCV 2001. vol. 2, pp. 426–432. Vancouver, USA (2001)
49. Xu, C., Cheng, L.: Efficient hand pose estimation from a single depth image. In: ICCV 2013. pp. 3456–3462. Sydney, Australia (2013)

http://www.springer.com/978-3-319-23436-6

Man-Machine Interactions 4
4th International Conference on Man-Machine Interactions, ICMMI 2015 Kocierz Pass, Poland, October 6-9, 2015
Gruca, A.; Brachman, A.; Czachórski, T.; Kozielski, S. (Eds.)
2015, XIX, 711 p. 228 illus., 80 illus. in color., Softcover
ISBN: 978-3-319-23436-6