Quantifying the effect of a colored glove in the 3D tracking of a human hand

Konstantinos Roditakis, Antonis A. Argyros

Institute of Computer Science - FORTH,
and
Computer Science Department - University of Crete
Heraklion, Crete, Greece

Abstract. Research in vision-based 3D hand tracking targets primarily the scenario in which a bare hand performs unconstrained motion in front of a camera system. Nevertheless, in several important application domains, augmenting the hand with color information so as to facilitate the tracking process constitutes an acceptable alternative. With this observation in mind, in this work we propose a modification of a state of the art method [12] for markerless 3D hand tracking, that takes advantage of the richer observations resulting from a colored glove. We do so by modifying the 3D hand model employed in the aforementioned hypothesize-and-test method as well as the objective function that is minimized in its optimization step. Quantitative and qualitative results obtained from a comparative evaluation of the baseline method to the proposed approach confirm that the latter achieves a remarkable increase in tracking accuracy and robustness and, at the same time, reduces drastically the associated computational costs.

1 Introduction

We are interested in the problem of tracking the 3D pose and full articulation of a human hand based on visual information acquired by an RGBD camera. The problem is interesting from a theoretical but also from a practical point of view, as its solution is valuable to a broad range of application domains. Given that human actions and intentions are manifested in the way hands move, a detailed and accurate estimation of this motion can support action interpretation and intention inference.

Clearly, the interest of the relevant research community is focused on the case of markerless tracking of the human hand(s). This is because markerless hand tracking is not invasive and poses far less restrictions to any application domain. Nevertheless, marker-based tracking is indeed useful and acceptable in many application domains. For example, in the domain of rehabilitation of patients suffering by stroke, hand motions are observed and quantified in a constrained laboratory setting. A scenario of a subject wearing a colored glove is not considered unacceptable, especially if in return, tracking accuracy and robustness is greatly improved. In other scenarios in the domain of wearable haptics research, the hand to be tracked is anyway augmented with devices of known form and appearance that could facilitate the hand observation process.

Motivated by the above observations, in this work we are interested in quantifying the level at which a state of the art 3D hand tracking method like [12] can benefit from

richer-than-markerless visual observations. To this end, we design a color glove and we modify the 3D hand model and objective function definition employed in [12] to enable the exploitation of the richer set of observations. Then, we perform extensive quantitative and qualitative experiments and a comparative evaluation of the proposed method with the baseline method of [12]. The obtained results demonstrate that the proposed approach achieves impressive performance/accuracy gains and justify fully our approach from a computer vision systems perspective. As an example, in a challenging sequence showing a normal-speed hand motion obtained at low (3 fps) frame rate, the proposed approach achieves half of the error compared to [12] by using only the $1/8$th of the computational resources.

1.1 Related Work

Markerless hand articulation tracking methods can be classified [3, 12] based on how candidate 3D hand poses are generated and tested against the observations. The appearance based approaches [1, 9, 15, 16, 21] generate a large set of hand configurations off-line. Visual features are extracted for each of the generated poses, resulting in a database where each pose is associated with image features. During online operation, comparable features are extracted from the acquired image(s) and searched for in the offline database. The reported solution is the stored pose that match the computed features. Model-based methods [2, 3, 6, 12–14, 18] generate hand poses, extract features and compare them to the observed ones at runtime. Typically, the 3D hand pose that best explains the available observations is estimated based on the solution of a high dimensional optimization problem.

Appearance-based methods are computationally more efficient compared to model-based ones, at the cost of having a fixed accuracy that depends on the density of sampling of the 3D hand pose space. Furthermore, they are more difficult to adapt to different problems, because changing the object to be tracked requires to generate the off-line database anew. In contrast, model-based methods can be adapted easily to different scenarios, since all that is required is a change of the model of the object to be tracked. Furthermore, accuracy improves as the computational budget increases. Despite the relatively high computational requirements of model-based methods, implementations that exploit GPGPUs have resulted in near real time performance [12]. Quite recently, a new approach to the problem has been proposed [10] based on quasi-random sampling of the parameter space.

While markerless tracking of bare hands is the more general formulation of the problem and, as such, the most interesting one, several works have dealt with the problem of marker-based tracking. In the case of some commercial systems, tracking relies on gloves augmented with sensors[1] or reflectors of infrared light[2]. Such systems provide accurate hand motion capture in real time, at the cost of expensive hardware that may obstruct the action of the hand. In order to alleviate these problems, researchers have tried to simplify the required hardware setup (both the glove design and the employed camera setup) by using color information. With the exception of [20], previous

[1] http://www.metamotion.com/hardware/motion-capture-hardware-gloves-Cybergloves.htm
[2] http://www.metamotion.com/motion-capture/optical-motion-capture-1.htm

work in color-based hand tracking dealt with limited application domains [19] or short sequences [5]. Wang and Popovic [20] proposed an appearance-based/data-driven 3D hand pose estimation method. Their approach relies on a distinctive color pattern appearing on the glove that provides unambiguous information on the pose of the hand and, thus, turns the hand pose estimation problem into a database lookup problem. While their method is not as accurate as traditional optical motion capture methods, it requires a single camera and an inexpensive cloth glove. A similar approach has been proposed by [17] to track humanoid robotic hands.

The approach proposed in this paper shares the motivation of the method of Wang and Popovic [20]. However, instead of using color information to cast the 3D hand tracking problem as a nearest neighbour database search problem, color information is used to drive a model-based, hypothesize and test approach. Thus, the accuracy of the proposed method is not bounded by the inevitable sparse sampling of hand poses from which appearance based methods suffer and can easily be adapted to deal with different problem variants (e.g., tracking two hands as in [13]).

2 Method Description

To track the full articulation of a hand wearing a color glove, we built upon the method presented in [12], where 3D hand tracking is formulated as an optimization problem on the 26D space of hand 3D pose and articulation parameters. A hand model consists of 37 appropriately transformed geometric primitives that are connected in a kinematic structure that matches closely that of a human hand. The optimization problem seeks for the 27D hand configuration[3] that minimizes the discrepancy between hand hypotheses and hand observations in visual input acquired by an RGBD camera. The optimization minimizes an objective function that compares a hand hypothesis to hand observations. The method in [12] deals with markerless observations of the human hand. In our case, a purposefully designed colored glove provides richer and less ambiguous visual input and is expected to improve the obtained tracking results. To do so, the 3D hand model and the objective function of the optimization problem need to be designed carefully so that tracking takes advantage of the additional information.

2.1 Glove design

The use of a color glove aims at facilitating the robust and accurate identification of specific parts of the hand. In this work, we aim to detect each finger and the palm as separate parts. In this direction, the palm is assigned with white color and the pinky, ring, middle, index and thumb fingers are assigned with the easily identifiable and discriminative red, blue, yellow, green, and pink colors, respectively (see top left thumbnail in Fig. 1). By following a model-based, hypothesize-and-test tracking method, different glove colors or patterns can be used depending on the application, without requiring time-consuming learning and training processes. For example, new colors should be learned and the

[3] The optimization space has one more dimension than the degrees of freedom of the hand model due to the quaternion representation of 3D hand orientation

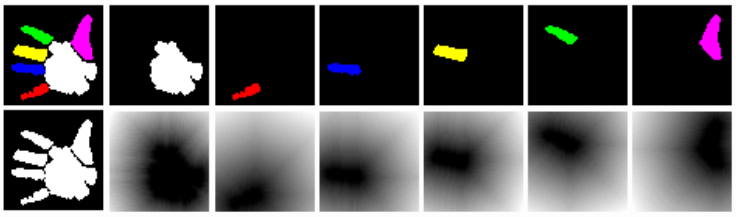


Fig. 1: Feature mapping of the segmented RGB glove image. 1st row: Segmented glove image and binary masks containing the individual hand parts. 2nd row: global foreground mask and distance transformed versions of the binary masks of the hand parts.

chromatic appearance of the 3D hand model needs to be specified. Still, this does not depend on the hand configuration, as happens with discriminative/appearance based methods that have to learn the appearance of hands in different poses anew.

2.2 Hand model

We adopt the kinematic and geometric structure as well as the parametrization of the hand model defined in [12]. The kinematic structure of the hand is defined as a 26-DOF forward kinematics model. Six DOFs represent its global position and orientation and 20 DOFs model the articulation of the five fingers. These 26 DOFs define the parametric space S on which optimization is performed. More precisely, they translate to a total of 27 parameters due to the redundant quaternion representation of hand orientation.

As [12] considers a skin colored hand, geometric primitives are not assigned specific colors. What is important in that work is the 3D structure of the hand in its various configurations. In contrast, in the 3D hand model employed in this work we define a coloring of the relevant primitives to match the glove design.

2.3 Preprocessing

The raw input consists of an RGB image I and an accompanying depth map D, both in VGA resolution. To filter-out noise, I is first smoothed with a Gaussian kernel. To perform color classification, we employ a standard method employing a color-based Gaussian Mixture Model (GMM) [4]. Given a set of training images of the glove and their manual segmentation, we build a GMM representation of each of the glove colors. Each color is represented as a mixture of three Gaussians. The parameters of the Gaussians and the mixture weights are estimated based on the EM algorithm. At run time, each pixel is assigned a color class label, depending on the probability that this color is drawn from each model GMM. We eliminate noisy labels from the segmented image by applying morphological filtering.

This segmentation is then employed to generate several features that are provided as observation input V to our method. First, a foreground binary mask F is generated. The pixels of F are set to 1 for glove pixels and 0 for non-glove pixels (see the bottom left image in Fig. 1). A distance map M is also generated by applying the distance transform

to F. M has a value of 0 in all points that correspond to foreground points of F. All other points of M (background points of F) have a value that is equal to their distance from the closest foreground point in F. From D and F a new depth map S is computed, where only depth values of D that correspond to glove pixels in F are kept. Maps F_l, M_l and S_l are computed for each and every individual color class l belonging to the set of color/label classes L. The observation input V consists of $V = \{F, S, F_l, M_l, S_l\}$.

2.4 Hand model rendering

The evaluation of a hand hypothesis h requires its comparison to the observation input V. To achieve this, h needs to be transformed to features comparable with observations. This is achieved by decoding the 27D hypothesis h through graphics rendering. To shape the hand, we apply forward kinematics over the parameters of h. Given camera calibration information C, a simulated labels map and a simulated depth map D' is acquired. Theses maps correspond to the labels map and D map that result from the acquisition and preprocessing steps of the RGB-D information. Having those, it is straightforward to compute all other relevant maps. Thus, comparable features R, which result from the rendering of a hypothesis h, consist of the set $R = \{F', S', F'_l, S'_l\}$.

2.5 Objective function and its optimization

During the optimization process, Particle Swarm Optimization (PSO) generates a set of candidate poses that need to be evaluated. An objective function E is defined that quantifies the discrepancy of a hypothesized hand pose h to the observed hand pose. Thus, the best-scoring hypothesis is the one that minimizes the objective function. The input to the objective function is the observation input features V (see Sec. 2.3) and a hand hypothesis h. As explained in Sec. 2.4, by assuming knowledge of the calibration parameters C of the employed RGBD sensor, we can transform the hand hypothesis h to features R that are comparable to observations V. The objective function $E(V,R)$ consists of the linear combination of three terms:

$$E(V,R) = w_1 K(V,R) + w_1 M(V,R) + w_2 DT(V,R). \tag{1}$$

In the above equation, $w_1 = 0.02$ and $w_2 = 0.002$ are experimentally determined weighting factors. The term $K(V,R)$ is defined as

$$K(V,R) = \frac{\sum min(|S - S'|, T)}{\sum (F \vee F') + \varepsilon} + \alpha \left(1 - \frac{2 \sum F \wedge F'}{\sum F \wedge F' + \sum F \vee F'}\right). \tag{2}$$

The term $K(V,R)$ serves two purposes, (a) penalizes depth discrepancies between the S and S' depth maps (left term) and (b) penalizes discrepancies between the F and the F' foreground masks (right term). In Eq.(2), $\alpha = 75$ is an experimentally determined normalization parameter, T is a clamp value set to $40mm$ and $\varepsilon = 10^{-6}$ is added to avoid possible divisions by zero. In the first term of $K(V,R)$, the sum of depth differences is clamped with the predetermined threshold T, in order to prevent the influence of noisy observations. Differences are normalized over their effective areas. In the second term of $K(V,R)$, $F \vee F'$ and $F \wedge F'$ yield binary maps, which are the result of the

pixel-wise disjunction and conjunction respectively, of the corresponding binary maps. Summations yield the number of foreground pixels in the resulting maps. Essentially, the second term of $K(V,R)$ represents the F1-score[4] between F and F'.

The term $M(V,R)$ is similar to $K(V,R)$, however, it considers all different hand parts (the palm and each of the five fingers) individually. For a particular label l, we define:

$$M(V,R,l) = \frac{\sum min(|S_l - S'_l|, T)}{\sum(F_l \vee F'_l) + \varepsilon} + \alpha\left(1 - \frac{2\sum F_l \wedge F'_l}{\sum F_l \wedge F'_l \; + \sum F_l \vee F'_l}\right). \tag{3}$$

Given the above definition of $M(V,R,l)$, $M(V,R)$ is then defined as

$$M(V,R) = \frac{1}{|L|}\sum_{\forall l \in L} M(V,R,l), \tag{4}$$

where $|L|$ is the number of all labels.

The last term $DT(V,R)$ of the objective function is again dependent on each label and takes into account the distance transformed versions of the observations. We define $DT(V,R,l)$ as

$$DT(V,R,l) = \frac{1}{N'_l}\sum M_l F'_l. \tag{5}$$

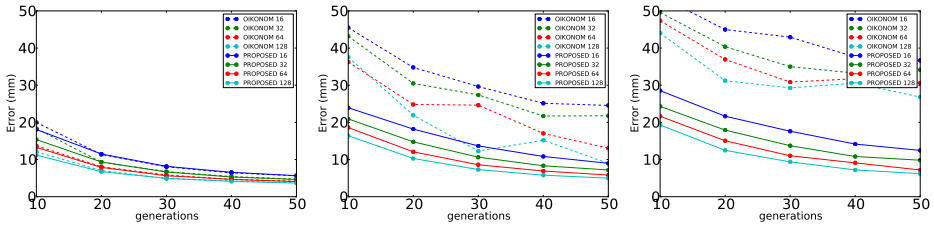
The pixelwise multiplication of M_l with F'_l is zero when the space occupied by each color label in the hypothesis is exactly the same as the space it occupies in the observation. Thus, the perfect hypothesis for the location of a hand part does not introduce any penalty in this term and the objective function. The summation is normalized with the number N'_l of image points of the rendered label l. Finally, $DT(V,R)$ is defined as

$$DT(V,R) = \frac{1}{|L_o|}\sum_{\forall l \in L_o} DT(V,R,l), \tag{6}$$

where L_o is the set of observed labels.

The optimization of the resulting objective function is performed based on Particle Swarm Optimization (PSO) [7, 8] as suggested in [12]. PSO is a population based stochastic optimization method that iteratively searches the optimal value of a defined objective function in a specified search space. The population (swarm) consists of particles each of which represents a candidate solution to the problem. The candidate solutions evolve in runs which are called generations, according to a policy which emulates social interaction. At the end, the candidate solution that achieved the best objective function score, through all generations, is selected as the estimated solution. The number of particles N and generations G determine the computational requirements of the optimization process, as their product defines the number of objective function evaluations. More details on the application of PSO to the problem of 3D hand tracking as well as information on the relevant tracking loop are reported in [12].

[4] `http://en.wikipedia.org/wiki/F1_score`



(a) M_D, 30 fps (b) M_D, 6 fps (c) M_D, 3 fps

Fig. 2: Median error M_D for hand pose estimation (in mm) for the proposed method (solid lines) is illustrated in comparison to that of [12] (dashed lines). In all plots the performance of different PSO budgets is shown. The synthetic sequence used represents the same hand motion sampled at(a) 30fps (b) 6 fps, (c) 3 fps.

3 Experimental Evaluation

Synthetic as well as real-world sequences were used to evaluate experimentally the performance of the proposed method. In all performed experiments, we compare quantitatively and qualitatively the proposed method with the baseline [12]. To do so we utilize the same C++ code base for both methods, a fact that facilitates their fair comparison.

The quantitative evaluation of the proposed method has been performed using synthetic data. This approach for evaluation is often used in the relevant literature [6, 11] because ground truth from real-world image sequences is hard to obtain. In order to synthesize realistic hand motions we first acquired a real-world sequence (30 fps, VGA resolution) with an Xtion sensor[5]. Then, we tracked this sequence using [12]. The resulting track was used to generate synthetic sequences through rendering.

To quantify the accuracy for a given optimization configuration, we adopt the metric used in [6]. For each processed frame, we compute the mean Euclidean 3D distance between the estimated phalanx endpoints and their ground truth counterparts. The average of this measure for all frames constitutes the resulting error estimate D. Since the optimization is stochastic, each distinct configuration was conducted 20 times. The median M_D of the resulting $D_{i=1..20}$ errors quantifies the accuracy obtained for a particular set of particle and generations counts.

In a first experiment, we employed a synthetic dataset of 682 frames where a hand performs motions and gestures, such as finger counting, closed fist formation, palm rotations, etc. For this dataset, two variants were identified. In the first, all hand parts were assumed to be uniformly colored, so that the sequence can be fed to the method of [12]. In the second variant, hand parts were colored according to our colored glove design, so as to give rise to input that is appropriate for the proposed approach.

Figure 2(a) shows the median errors M_d for both methods, as a function of PSO parameters. Each distinct line represents a particle configuration. We observe that the proposed work marginally outperforms [12], except for results coming from low PSO budgets. Both approaches exhibit similar behaviour as the particle and generation counts

[5] `http://www.asus.com/Multimedia/Xtion_PRO/overview`

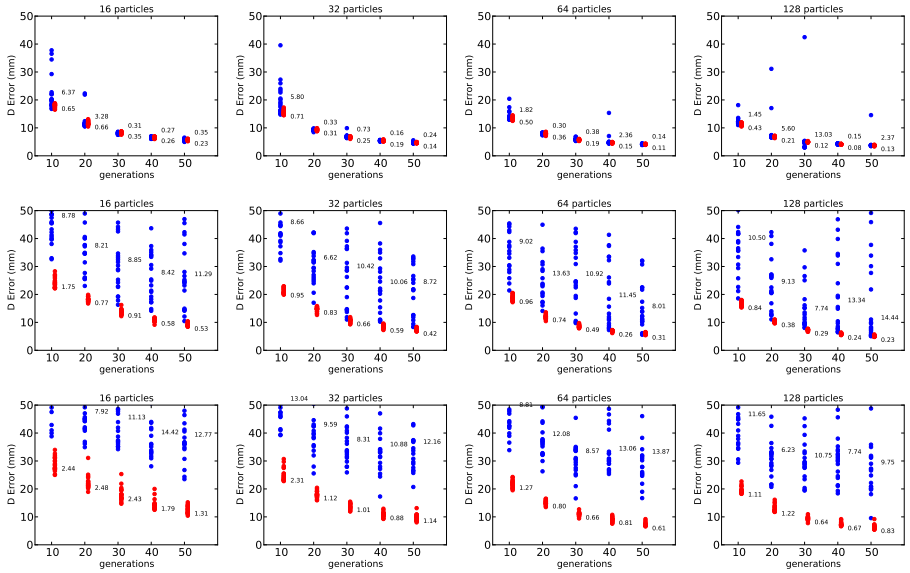


Fig. 3: Errors in hand tracking (in mm) for the proposed method (red points) in comparison to that of [12] (blue points). Each row corresponds to a different image acquisition fps. Figure columns correspond to different particle counts. Each point represents a single run. For a particular parameter set, 20 runs were performed. The number assigned to each group of points denotes the standard deviation of the error is these 20 runs.

increase. The increase of accuracy for configurations larger than 32 particles and 30 generations is very small compared to to the increase of the computational budget.

Individual errors D take values between $3.3mm$ and $18.7mm$ for the proposed method and between $2.9mm$ and $52.3mm$ for [12]. Moreover, as shown in the first row of plots of Fig 3, the resulting standard deviation for low computational budgets ($N = 16$, $N = 32$ and $G = 10$), is an order of magnitude lower for the proposed method ($\sigma \simeq 0.7mm$) compared to that of [12] ($\sigma \simeq 6mm$).

We use the same dataset but we only employ one out of each 5 frames. The resulting synthetic sequence is subject to a dual interpretation. It is either a dataset which contains a hand acting five times faster than the previous one, or a dataset with five times lower acquisition rate. Lower acquisition rates occurs either because of higher computational requirements, or when the framework is employed on weaker hardware. Moreover, in the new dataset a decrease in accuracy is expected due to larger displacement of the hands between consecutive frames.

Figure 2(b) shows the corresponding median errors obtained in the new dataset. It can be verified that the proposed method performs strikingly better compared to [12]. The performance of our method for 16 particles is better than the performance of [12] for 128 particles. This constitutes a more than 8-fold reduction of the computational time. Additionally, in the second row of Fig. 3, the obtained errors D are visualized

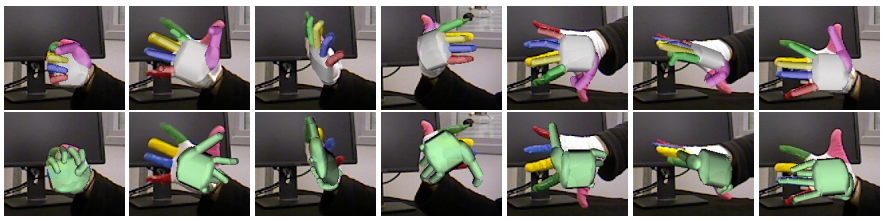


Fig. 4: Snapshots from an experiment with real data. Cropped 320×240 regions from the original 640×480 images are shown. 1st row: proposed method, 2nd row: the performance of the method in [12] with the same computational budget (64 particles running for 50 generations). See text for details.

for all 20 experiments and for different particle counts. As it can be verified, in our method errors are not only smaller in absolute terms, but the variability in performance in different runs is also very much decreased.

The third experiment is similar to experiment 2, with the difference that we temporally subsample the original dataset every 10 frames (3 fps). Figure 2(c) as well as the third row of Fig. 3 demonstrate that the difference in performance between the proposed and the baseline method is further widened, in favour of the proposed method.

Towards the qualitative evaluation of the proposed approach in real data, several long real-world image sequences were captured. Figure 4 illustrates indicative results obtained from one such sequence (690 frames) where the human hand wears the designed color glove. As it can be observed in the first row of snapshots, the estimated hand models are in very close agreement with the image data, despite the complex articulation of the performing hand and the low (3 fps) framerate. The color information was also processed to come up with a full hand segmentation so as to make the method in [12] applicable to this input. As shown in Fig. 4, 2nd row, [12] fails in this case. A video accompanying the paper[6] illustrates the obtained results.

4 Summary

The interest on markerless 3D hand tracking is unquestionable. Nevertheless, the interest in systems that perform mildly invasive, marker-based, 3D hand tracking is also undeniable. In this work we capitalized on a powerful 3D hand tracking approach and we enabled it to benefit from color information that disambiguates hand parts. We achieved this by purposefully designing a color glove and by appropriately modifying the hand model and the objective function of the baseline approach. Quantitative and qualitative experimental results demonstrate that our approach achieves dramatic accuracy improvement over the baseline (smaller errors and error variances in several runs) with a fraction of the computational resources. Future work includes the design and the evaluation of gloves tailored to specific applications, in order to minimize invasiveness and maximize accuracy, robustness and computational performance.

[6] http://youtu.be/9nkHIgFYtzE

Acknowledgements

This work was partially supported by the EU FP7-ICT-2011-9 project WEARHAP.

References

1. V. Athitsos and S. Sclaroff. Estimating 3D Hand Pose From a Cluttered Image. In *CVPR*, pages II–432–9. IEEE, 2003.
2. Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *ECCV*, pages 640–653, 2012.
3. M. de La Gorce, D.J. Fleet, and N. Paragios. Model-based 3d hand pose estimation from monocular video. *IEEE Trans. on PAMI*, pages 1–15, February 2011.
4. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.
5. B. Dorner. *Chasing the colour glove: Visual hand tracking*. PhD thesis, Simon Fraser University, 1994.
6. H. Hamer, K. Schindler, E. Koller-Meier, and L. Van Gool. Tracking a Hand Manipulating an Object. In *ICCV*, 2009.
7. J. Kennedy and R. Eberhart. Particle Swarm Optimization. In *International Conference on Neural Networks*, volume 4, pages 1942–1948. IEEE, January 1995.
8. J. Kennedy, R. Eberhart, and S. Yuhui. *Swarm intelligence*. Morgan Kaufmann, 2001.
9. C. Keskin, F. Kirac, Y. Kara, and L. Akarun. Real time hand pose estimation using depth sensors. *ICCV Workshop*, pages 1228–1234, 2011.
10. I. Oikonomidis and A.A. Argyros. Evolutionary Quasi-random Search for Hand Articulations Tracking. In *IEEE CVPR*, Columbus, Ohio, USA, July 2014.
11. I. Oikonomidis, N. Kyriazis, and A.A. Argyros. Markerless and Efficient 26-DOF Hand Pose Recovery. *ACCV*, pages 744–757, 2010.
12. I. Oikonomidis, N. Kyriazis, and A.A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, Dundee, UK, Aug. 2011.
13. I. Oikonomidis, N. Kyriazis, and A.A. Argyros. Tracking the articulated motion of two strongly interacting hands. In *CVPR*, pages 1862–1869. IEEE, June 2012.
14. J.M. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *ICCV*, page 612, Los Alamitos, CA, USA, 1995. IEEE Computer Society.
15. J. Romero, H. Kjellstrom, and D. Kragic. Monocular real-time 3d articulated hand pose estimation. In *IEEE-RAS Int'l Conf. on Humanoid Robots*. IEEE, Dec 2009.
16. R. Rosales, V. Athitsos, L. Sigal, and S. Sclaroff. 3D hand pose reconstruction using specialized mappings. *ICCV*, 2001.
17. M. Schroder, C. Elbrechter, J. Maycock, R. Haschke, M. Botsch, and H. Ritter. Real-time hand tracking with a color glove for the actuation of anthropomorphic robot hands. In *2012 12th IEEE-RAS International Conference on Humanoid Robots*, pages 262–269. IEEE, 2012.
18. EB Sudderth, MI Mandel, WT Freeman, and AS Willsky. Visual hand tracking using nonparametric belief propagation. In *CVPR Wkshp on Generative Model-based Vision*, 2004.
19. C. Theobalt, I. Albrecht, J. Haber, M. Magnor, and H.-P. Seidel. Pitching a Baseball – Tracking High-Speed Motion with Multi-Exposure Images. In *Proc. SIGGRAPH '04*, pages 540–547. ACM SIGGRAPH, 2004.
20. R.Y. Wang and J. Popović. Real-time hand-tracking with a color glove. *ACM Transactions on Graphics*, 28(3), 2009.
21. Y. Wu and T.S. Huang. View-independent recognition of hand postures. In *CVPR*, volume 2, pages 88–94. IEEE, 2000.