Full DOF Tracking of a Hand Interacting with an Object by Modeling Occlusions and Physical Constraints

Iason Oikonomidis, Nikolaos Kyriazis, Antonis A. Argyros
Institute of Computer Science - FORTH and Comp. Science Department, Univ. of Crete
Heraklion, Crete, Greece
{oikonom,kyriazis,argyros}@ics.forth.gr

Abstract

Due to occlusions, the estimation of the full pose of a human hand interacting with an object is much more challenging than pose recovery of a hand observed in isolation. In this work we formulate an optimization problem whose solution is the 26-DOF hand pose together with the pose and model parameters of the manipulated object. Optimization seeks for the joint hand-object model that (a) best explains the incompleteness of observations resulting from occlusions due to hand-object interaction and (b) is physically plausible in the sense that the hand does not share the same physical space with the object. The proposed method is the first that solves efficiently the continuous, full-DOF, joint hand-object tracking problem based solely on markerless multicamera input. Additionally, it is the first to demonstrate how hand-object interaction can be exploited as a context that facilitates hand pose estimation, instead of being considered as a complicating factor. Extensive quantitative and qualitative experiments with simulated and real world image sequences as well as a comparative evaluation with a state-of-the-art method for pose estimation of isolated hands, support the above findings.

1. Introduction

The estimation of the full pose of hands from markerless visual observations is a problem whose solution is of fundamental importance in numerous applications including but not limited to the visual perception of grasping and manipulation, sign language understanding, human computer interaction, etc. It is also known that a number of cascading issues such as the dimensionality of the problem, the incomplete and/or ambiguous observations due to scene clutter and occlusions as well as the requirement for accurate estimates in real time, hinder its effective solution.

Full DOF hand pose recovery during hand-object interaction is a much more interesting problem but also more difficult due to the induced hand-object occlusions. A common approach is to treat the object as a "distractor" that nevertheless leaves some partial evidence which is enough for hand pose estimation. On the contrary, our goal is to see the manipulated object as a source of useful constraints.

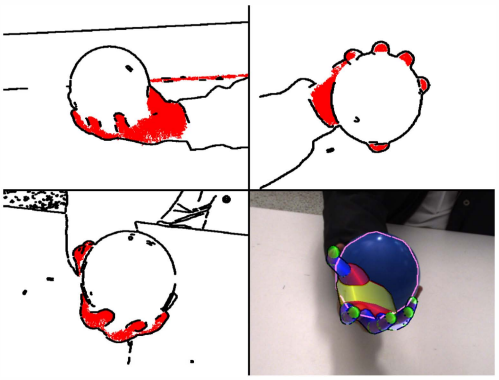


Figure 1. Top row, and bottom left: Three views of a hand grasping an object. Skin regions appear in red and edges in black. The hand is partially occluded by the object in all views. The incomplete skin and edge maps of the hand facilitate the generation of a hypothesis for a hand manipulating a compact sphere. At the same time, given this hypothesis, the 3D pose of the hand can be estimated more accurately. Bottom right: the output of the proposed approach superimposed in one of the frames.

In this work, it is assumed that hand-object interaction is observed by a multicamera system. In each of the acquired views, edge and skin color maps form 2D cues of the presence of a hand. Depending on the viewpoint, the presence of an object occludes the performing hand (Fig. 1). This incomplete observation of the hand provides important evidence on the type and pose of the manipulated object. Conversely, attributing missing hand observations (skin color, edges) to the presence of a manipulated object permits a more accurate estimation of the pose of the partially observed hand. Another source of useful constraints stems from the properties of the natural world, i.e., the fact that the hand and the object cannot share the same physical

2011 IEEE International Conference on Computer Vision
978-1-4577-1102-2/11/$26.00

space. Thus, the 3D shape and pose of the object provides important information on the articulation of the hand and vice versa. The tight coupling between "what the hand tells about the object" and "what the object tells about the hand", suggests that we should identify *simultaneously* the hand configuration and the object 3D model and pose that best explain the observed scene holistically. In this spirit, we formulate an optimization problem that takes into account the constraints mentioned above. Thus, the solution to this problem is a joint hand-object model that besides being compatible to the available visual observations, is also physically plausible.

1.1. Related work

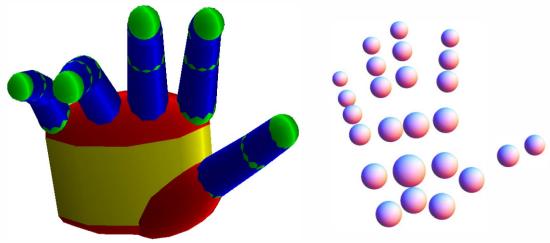
The recovery of the full 3D configuration of articulated objects such as humans and hands presents a lot of challenges. Several approaches have been proposed that address various aspects of the problem such as its dimensionality, the incomplete and/or ambiguous observations due to scene clutter, its computational requirements, etc. Moeslund et al. [15] provide a review of research to the general problem of visual human motion capture and analysis. A review that is specific to the problem of human hand motion estimation is provided in [7]. Related methods are categorized as partial or full pose estimation methods, depending on the level of detail they provide regarding the observed hand.

Another categorization identifies appearance-based and model-based methods. Appearance-based methods estimate hand configurations by establishing a direct mapping of image features to the hand configuration space [3, 22, 28, 20]. Model-based approaches employ a 2D or a 3D hand model [19, 25, 26, 6, 16]. In the case of 3D hand models, the hand pose is estimated by matching the projection of the model to the observed image features. The task is then formulated as a search problem in a high dimensional configuration space, which typically induces a high computational cost. In our work presented in [17], we proposed an efficient (15 fps) method for tracking the articulation of a hand, which depends on visual input provided by the Kinect sensor [14]. A common characteristic of all the methods mentioned above is that they consider human hands in isolation. Thus, in the context of hand-object interaction, their accuracy in hand pose estimation is compromised due to the induced hand-object occlusions that affect drastically the completeness of hand observations.

Given the significant role of context in human visual recognition [18], several researchers have attempted to exploit contextual constraints in solving computer vision problems. Closely related to our problem, a few recent works [12, 8, 29, 21] consider context for classifying human-object interaction activities. The related methods can be classified based on whether they refer to the human body or hand and also on whether they provide a detailed 3D model of the actor (human body or hand) and the object. Thus, [8, 29] study the full human body while in interaction with objects. From these, only [29] provides detailed information on human body pose. For the same problem, Gupta et al. [9] and Sigal et al. [23] propose solutions to handling self-occlusions but not occlusions with other objects. Kjellstrom et al. [12] consider hand-object interactions but only for classifying them, without providing a detailed hand and object model. The work of Hamer et al. [10] also addresses the problem of hand-object interaction but depends on the use of a structured light range sensor and does not model the manipulated object. Finally, Romero et al. [21] propose a method for estimating the pose of a hand interacting with an object which is appearance-based. A method that exploits context to provide a detailed 3D model for both hands and objects is missing from the current literature. The proposed method is trying to fill this gap.

Towards this direction, in this work we extend the approach in [16] by considering jointly the hand and the manipulated object. In [16], a generative, multiview method for 3D hand pose recovery is presented. In each of the acquired views, reference features are computed based on skin color and edges. A 26-DOF 3D hand model is adopted. For a given hand configuration, skin and edge feature maps are rendered and compared directly to the respective observations. The discrepancy of a given 3D hand pose to the observations is quantified by an objective function that is minimized through Particle Swarm Optimization (PSO). The whole approach is implemented efficiently on a GPU. In the proposed approach, we do not only seek for the optimal hand model that explains the available hand observations but rather the joint hand-object model that best explains both the available hand/object observations and the occlusions. Additionally, the objective function penalizes hand-object penetration, seeking for a physically plausible solution. It is demonstrated that the aforementioned constraints are very useful towards an accurate solution to this more complex and interesting problem.

The proposed approach is the first model-based method that solves efficiently, the continuous, full-DOF, joint hand-object tracking problem based on markerless camera input. Additionally, it is the first to demonstrate that hand-object interaction is not necessarily a complicating factor towards estimating the configuration of a hand but a context that can be exploited effectively towards a more accurate solution. As an additional result, the method provides a parametric 3D model of the manipulated object together with its 3D position and orientation. This is achieved by exploiting the hand-object occlusions and despite the fact that only a parametric representation of the object's 3D shape is known that is lacking an explicit appearance model. The approach explores an infinite configuration space. Thus, its accuracy is not limited by the size and content of a database of hand



(a) (b)

Figure 2. Graphical illustration of the employed 26-DOF 3D hand model, consisting of 37 geometric primitives (a) and the 25 spheres constituting the hand's collision model (b).

configurations, as e.g. in [20]. The above are supported by qualitative and quantitative experiments with both simulated and real world image sequences as well as by a comparative evaluation with the method in [16].

2. Hand-object pose estimation (*HOPE*)

The problem of joint hand-object pose estimation is formulated as a multidimensional optimization problem. In the following, we present in detail the basic building blocks of the proposed method for joint Hand-Object Pose Estimation (*HOPE*), with emphasis on the employed observation model, joint hand-object 3D model, hypothesis evaluation mechanism and optimization method.

2.1. Computed visual cues

The proposed method operates on sequences of synchronized views acquired by intrinsically and extrinsically calibrated cameras. A set of images acquired from these cameras at the same moment in time is called a *multiframe*. If $M_i = \{I_1, I_2, \ldots\}$ is a multiframe of a sequence $S = \{M_1, M_2, \ldots\}$, I_j denotes the image from the j-th camera/view at the i-th time step. For each image I of a multiframe M, an edge map $o_e(I)$ is computed through Canny edge detection [4] and a skin color map $o_s(I)$ is computed using the method presented in [2]. As a convention, 1 indicates presence and 0 indicates absence of skin or edges in the respective maps. For each edge map $o_e(I)$, a distance transform $o_d(I)$ is computed. Maps $\{o_s(I), o_d(I)\}$ constitute the observation cues for image I.

2.2. Joint hand-object model

The proposed approach employs a model $m = (h, o)$ that represents jointly a hand h and the manipulated object o. The hand model h consists of a palm and five fingers. The palm is modeled as an elliptic cylinder and two ellipsoids for caps. Each finger consists of three cones and four spheres, except for the thumb that consists of two cones, an ellipsoid and three spheres (Fig. 2, (a)). 25 additional spheres constitute the hand's collision model and are used for checking for hand-hand and hand-object interpenetration (Fig. 2, (b) and Sec. 2.3). The resulting 62 3D geometric primitives of the hand model are different parameterizations of an ellipsoid and a truncated cylinder. The assembly of appropriate homogeneous transformations of these two geometric primitives yields a hand model similar to that of [25].

The kinematics of each finger is modeled using four parameters encoding angles, two for the base of the finger and two for the remaining joints. Ranges of parameter values are determined based on anatomical studies [1]. The position of a fixed point on the palm defines the global position of the hand. The global orientation is parameterized using the redundant representation of quaternions. This parameterization results in a 26-DOF model encoded in a vector of 27 parameters.

For representing an object, in principle, any parametric model o can be used. The representation of common handheld objects such as cuboids, ellipsoids and cylinders requires 3, 3 and 2 intrinsic shape parameters, respectively. More complex parametric shape models like superquadrics require as many as 6 parameters. Regardless of the intrinsic shape parameterization, 7 additional parameters are required, 3 for 3D position and 4 for a quaternion-based representation of 3D orientation. In this work, we provide experimental results with ellipsoids, cuboids and cylinders. Nevertheless, there is no inherent limitation that prevents the method from being able to handle more complex object models, provided that this does not increase the dimensionality of the problem prohibitively. Interestingly, a complex, known 3D object represented as a mesh has less DOFs compared to our parametric object models (6 DOFs, 3D pose).

2.3. Evaluation of hand-object model hypotheses

Given a joint parametric hand-object model $m = (h, o)$, the goal is to estimate the parameters that give rise to the hand-object configuration that (a) is most compatible to the image features present in multiframe M (Sec. 2.1) and (b) is physically plausible in the sense that two different rigid bodies cannot share the same physical space (interpenetration constraints). To achieve this, an objective function $O(m, M)$ is defined as:

$$O(m, M) = \sum_{I \in M} D(I, m) + \lambda_k W(m). \qquad (1)$$

In Eq.(1), the first term quantifies the discrepancies of a given hand-object model m to the actual camera-based observations, while the second term quantifies the penetration depth between the hand and the object, but also among hand parts (fingers, palm, etc). λ_k is a weighting factor experimentally set to $\lambda_k = 0.1$.

To compute $D(I, m)$, we first compute comparable image features from each hypothesized hand-object model.

More specifically, an edge map $r_e(m)$ and a skin color map $r_s(m)$ can be generated my means of rendering. The reference implementation of the rendering process is similar to that of [25]. The implicit assumption made at this point is that an object cannot contain skin-colored pixels. Thus, the hand component h of m contributes to the skin color map $r_s(m)$ by setting visible hand pixels to 1, while the object component o of m contributes to the skin color map $r_s(m)$ by setting map pixels to 0. Experimental results have verified that the presence of a moderate number of skin-colored pixels on the object's surface does not affect the accuracy of the method. $D(I,m)$ is then defined as:

$$D(I,m) = 1 - \frac{2\sum o_s(I) \wedge r_s(m)}{(\sum o_s(I) \wedge r_s(m)) + (\sum o_s(I) \vee r_s(m))} + \lambda \frac{\sum o_d(I) \cdot r_e(m)}{\sum r_e(m) + \epsilon}, \quad (2)$$

where $o_s(I), o_d(I)$ are defined in Sec. 2.1 and ϵ is a small quantity to prevent division by zero. The 1st row of Eq.(2) models the discrepancies between the skin-colored pixels of the model and the observations. Sums are computed over entire feature maps. In contrast to [16], this part of the objective function is normalized to the interval $[0..1]$. The 2nd row models the discrepancies between the rendered edge maps and the observed edge maps. This is achieved by summing the values of the distance-transformed observation edge map that concur with the edges of the rendered model. λ is a constant normalization factor that was set to 0.02 in all experiments.

The role of function $W(m)$ in Eq.(1) is to penalize (a) hand configurations where hand parts intersect each other (self-penetration) and (b) hand-object configurations where the hand h intersects the object o (interpenetration). Let $P(p_i, p_j)$ be the minimum magnitude 3D translation that is required so that the volume of intersection of geometric primitives p_i and p_j becomes equal to 0. This is effectively computed using the *Open Dynamics Engine* (ODE) [24]. Let also S_h be the primitives of the hand's collision model, as shown in Fig.2(b). The self-penetration P_{hh} of a given hand configuration is defined as $P_{hh} = \max_{i \in S_h, j \in S_h, i \neq j}\{P(i,j)\}$. The interpenetration P_{ho} is similarly defined as $P_{ho} = \max_{i \in S_h}\{P(i,o)\}$. Then, $W(m)$ is defined as $W(m) = \max\{P_{hh}, P_{ho}\}$. Thus, both self- and inter- penetrations are treated in a uniform manner. Additionally, self-penetration is treated more systematically compared to [16] where only certain abduction-adduction angles between adjacent fingers were penalized.

2.4. Optimization

The minimization of the objective function of Eq.(1) is achieved through PSO. Introduced by Kennedy et al. [11], PSO achieves optimization through a policy which emulates the "social interaction" of a population of atoms (particles) that evolves in a number of generations. A population is essentially a set of particles that lie in the parameter space of the objective function to be optimized.

Following the notation introduced in [27], every particle holds its current position (current candidate solution, set of parameters) in a vector x_t and its current velocity in a vector v_t. The ith particle stores in vector p_i the position which corresponds to the best evaluation of its objective function up to the current generation t. All particles of the swarm become aware of the current global optimum p_g, the best position encountered across all particles of the swarm. In every generation t, the velocity of each particle is updated according to $v_t = K(v_{t-1} + c_1 r_1 (p_i - x_{t-1}) + c_2 r_2 (p_g - x_{t-1}))$ and its position according to $x_t = x_{t-1} + v_t$. In the above equations, K is a constant constriction factor [5], c_1 is called the cognitive component, c_2 is termed the social component and r_1, r_2 are random samples of a uniform distribution in $[0..1]$. Finally, $c_1 + c_2 > 4$ must hold [5]. In all performed experiments the values $c_1 = 2.8$, $c_2 = 1.3$ and $K = 2/|2 - \psi - \sqrt{\psi^2 - 4\psi}|$ with $\psi = c_1 + c_2$ were used.

The search space is a multidimensional cuboid. The particle positions are initialized randomly and the particle velocities are set to zero. If, during the position update, a velocity component forces the particle to move outside the search space, this component is zeroed and the particle does not perform any move at the corresponding dimension. The final outcome of the PSO is p_*, the particle with the best score across all generations.

The search space of *HOPE* is the joint hand-object model parameter space m. Given a hand model represented by 27 parameters and an object model represented by d parameters, the search space has $(27 + d)$ dimensions. The objective function to be minimized is $O(m, M)$ and the population is a set of hypothesized 3D hand-object configurations. The outcome of PSO $p_* = m_* = \text{argmin}_m(O(m, M))$ represents the best guess of the algorithm for the joint hand-object model parameters m given the multiframe M.

In the first multiframe, tracking is manually initialized by placing the hand roughly at a predefined position and pose. We have verified experimentally that model estimation is successful with discrepancies of up to $10cm$ in position and up to $20deg$ in global hand pose. Finger joints are correctly estimated if the above constraints are met. To cope with the tracking of the hand-object configuration in time, temporal continuity is exploited. The solution for multiframe M_{t-1} is used to bootstrap the initial population for the optimization problem of M_t. The first member of the population m_{ref} for M_t is the solution for M_{t-1}; The rest of the population consists of perturbations of this solution. The optimization for multiframe M_t is executed for a fixed amount of generations/iterations. After all generations have evolved, the best hypothesis m_* is dubbed as the solution for time step t.

3. Experimental Evaluation

The proposed method has been validated extensively based on both synthetic and real-world sequences of multiframes. First, we demonstrate the accuracy and the computational performance of the proposed (*HOPE*) method on a synthetically rendered data set where hands perform different grasps on a variety of objects (Sec. 3.1). We also compare the performance of *HOPE* to that of the method in [16], hereafter abbreviated as *PEHI* (Pose Estimation of Hands in Isolation). A final experiment with synthetic data involves the application of *HOPE* to a data set showing hands in isolation. The goal of this experiment is to show that *HOPE* can also estimate the pose of hands in isolation effectively, as a special case.

Besides the synthetic data, we also provide qualitative evidence on how the *HOPE* and *PEHI* perform on real sequences of multiframes (Sec. 3.2). Although ground truth information is not available, these indicative results confirm the superiority of *HOPE* over *PEHI* which is in accordance with the experimental results over synthetic data.

3.1. Experiments on synthetic data

Experiments with synthetically produced sequences of multiframes were performed to enable the assessment of the proposed method based on ground truth data. To that end, we simulated different grasps of three different objects (an ellipsoid, a cylinder, and a box) performed by the employed hand model (Sec. 2.2). The interaction of the hand with each of these three objects was observed by 8 virtual cameras surrounding the scene. This resulted in three sequences consisting of 116 multiframes of 8 frames, each. The required cue maps (edges, skin color) were synthesized through rendering (Sec. 2.2).

For the quantitative evaluation of the method, an error metric quantifying the discrepancy between a true hand pose and an estimated hand pose is required. This was computed as follows. The five fingertips as well as the center of the palm were selected as reference points. For each such reference point, the Euclidean distance between its estimated position and its ground truth position is first calculated. These distances are averaged across all multiframes of each sequence, and all sequences. This results in a single error value $\mathcal{D}$ for the whole dataset.

Figures 3(a) and (c) illustrate the estimated error $\mathcal{D}$ of the *HOPE* method as a function of the PSO parameters. In Fig. 3(a), $\mathcal{D}$ is plotted as a function of the number of PSO generations and particles per generation, for multiframes consisting of 2 views. The cameras providing these two views are placed opposite to each other. $\mathcal{D}$ takes values between $22mm$ and $55mm$. It can be verified that for more than 30 generations and more than 32 particles/generation the error in 3D hand pose recovery for *HOPE* does not vary considerably and it is in the order of $25mm$.

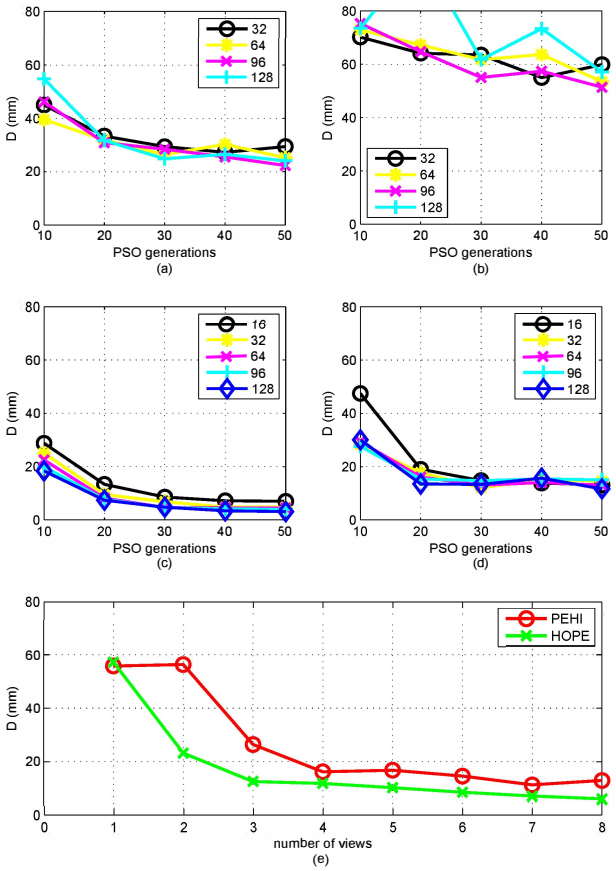


Figure 3. Mean error $\mathcal{D}$ for hand pose estimation (in mm) for *HOPE* (left) and *PEHI* (right) for different PSO parameters and number of views. (a), (b): Varying PSO particles and generations for 2 views. (c),(d): Same as (a),(b) for 8 views. (e): Increasing number of views, 40 generations, 64 particles/generation.

Figure 3(b) is analogous to that of Fig. 3(a) for the *PEHI* algorithm. In this case, the mean error $\mathcal{D}$ does not decrease monotonically as a function of particles. This is attributed to the incomplete/occluded hand observations that undermine the convergence of *PEHI*. $\mathcal{D}$ now ranges between $51mm$ and $101mm$. It can be verified that for more than 30 generations and more than 32 particles/generation the error in 3D hand pose recovery for *PEHI* is in the order of $55mm$. Thus, the error of *PEHI* is on average more than twice the error of *HOPE*.

Figures 3(c) and (d) are analogous to Figs. 3(a) and (b), except the fact that each multiframe now consists of 8 rather than 2 views. $\mathcal{D}$ takes values between $3mm$ and $29mm$ for *HOPE* and between $12mm$ and $47mm$ for *PEHI*. For more that 30 PSO generations and more than 32 particles per generation the error of *PEHI* is still more than twice the error of *HOPE*. Interestingly, whatever *HOPE* achieves with 16 particles and 20 generations is equal or better to what *PEHI* achieves with any of the tested particles/generations combi-

Table 1. Estimated/actual parameters for the object models in the experiments with synthetic data.

Object	Estimated/Actual parameters (in mm)
Cylinder	Radius: 54/55, Height: 127/128
Ellipsoid	X: 54/55, Y: 83/85, Z: 126/128
Box	X: 77/77, Y 128/129, Z: 155/156

nations.

In order to better assess the behavior of the method with respect to the number of available views, additional experiments with a varying number of views were conducted. Figure 3(e) shows the behavior of *HOPE* and *PEHI* as a function of the size of a multiframe. For the experiments with less than 8 views, these were selected empirically to be as complementary as possible. PSO involved 64 particles running for 40 generations. The obtained results demonstrate that modeling the occluder and the physical constraints is more beneficial than adding an extra camera. As an example, exploiting these constraints with two cameras is still better than with three cameras and the hand alone. In fact, whatever *HOPE* achieves with three views is already better to what *PEHI* achieves with as many as eight.

Overall, the experiments in Fig. 3 show a consistent and significant superiority of *HOPE* over *PEHI* which is dominant in the case of a limited number of available views. This is important because it allows for practical joint hand-pose estimation by a multicamera system with a few cameras that is associated with less costs, complexity and requirements for computational resources.

Besides its superiority in hand pose estimation, *HOPE* also estimates the model parameters of the manipulated object. The average positional error of object detection across all sequences of multiframes in the experiments of Fig. 3 is $3mm$ (Euclidean distance between true and estimated positions) and the average orientation error is $2\,\mathrm{deg}$. Table 1 shows the actual and estimated object parameters. The later are averaged for all the multiframes of the sequence that depicts the corresponding object. It can be verified that for all types of objects, the estimated model parameters are very close to the ground truth.

The runtime[1] of a GPU-powered implementation of *HOPE* [13] for runs of 40 PSO generations and 64 particles per generation is 0.31sec for a single-view multiframe and 2.19sec for an 8-view multiframe. An online version of the system employing 4 cameras, operates at 2 fps.

In multiframes of sizes larger than 2, *PEHI* is approximately 20% faster than *HOPE*. This overhead is attributed to the computation of the $W(m)$ component of the objective function. Since this is a fixed overhead that is independent

[1]Experiments run on the computational infrastructure presented in Sec.3.2.

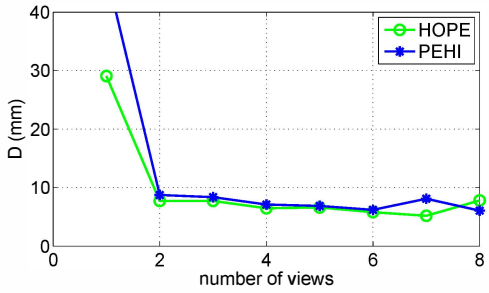


Figure 4. Performance of *HOPE* and *PEHI* on a synthetic sequence of multiframes that shows hands in isolation. 64 PSO particles and 40 generations have been used in both cases.

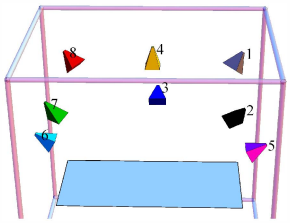


Figure 5. Camera setup for the experiments with real data.

of the multiframe size, the relative difference in computational performance decreases with the number of views.

Finally, we applied both *HOPE* and *PEHI* to a synthetic image sequence (400 multiframes, 8 frames/multiframe) showing non-rigid motion of hands in isolation. Figure 4 plots the mean error $\mathcal{D}$ as a function of the number of the employed views. For both algorithms, 40 PSO generations and 64 particles per generation were used. For *HOPE*, a cylindrical object has been hypothesized. The result shows that the performance of the two algorithms is comparable, a fact that indicates the capability of *HOPE* to track hands observed in isolation. Expectedly, *HOPE* estimated the presence of very small objects (size in the order of a few mms).

3.2. Experiments on real image data

Real-world image sequences were acquired using a multicamera system (Fig. 5) installed around a $2 \times 1m^2$ bench and consisting of 8 synchronized and calibrated *Flea2* PointGrey cameras. Each camera has a maximum framerate of 30 fps, at 1280×960 image resolution. However, the core processing is performed on 256×256 windows centered around the previous multiframe solution. The workstation for image acquisition and processing is equipped with a quad-core Intel i7 920 CPU, 6 GBs RAM and a 1581 $GFlops$ Nvidia GTX 580 GPU with 1.5 GBs RAM.

Three sequences of multiframes have been acquired, each showing a hand grasping and manipulating a spherical (301 multiframes), a cylindrical (261 multiframes), and a box (251 multiframes) object. Figure 6(a) provides sample results obtained by applying *HOPE* (top row) and *PEHI*

Table 2. Estimated/actual parameters for the object models in the experiments of Fig. 6.

Object	Estimated/actual parameters (in mm)
Cylinder	Radius: 51/53, Height: 121/131
Ellipsoid	X: 128/116, Y: 128/116, Z: 122/116
Box	X: 66/67, Y: 158/150, Z: 84/93

Table 3. The mean value of the objective function of *HOPE* and its standard deviation when optimization searches for cylinders, ellipsoids and cuboids for a sequence showing an ellipsoid (sphere).

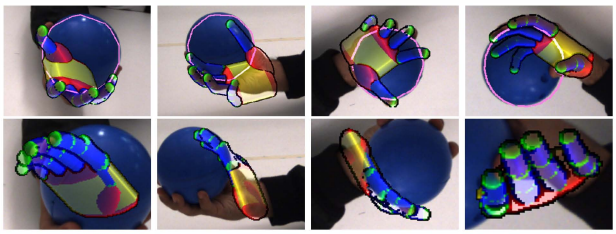
	Cylinder	Ellipsoid	Cuboid
Mean value	3.02	2.65	3.95
Stdev.	0.68	0.57	1.17

(bottom row) to a specific multiframe of the sphere sequence. Since the hand is mostly occluded by the sphere in all views, *HOPE* estimates the hand configuration correctly while *PEHI* fails completely. Similar results were obtained in the case of the cylinder sequence which shows a hand grasping and turning a cylindrical object up-side down. Fig. 6(b) shows four frames acquired from the same camera in different moments in time. *HOPE* tracks the configuration of the hand throughout the whole sequence whereas *PEHI* looses track of the hand as soon as the later becomes severely occluded by the object. Figure 6(c) shows a similar result for the box sequence. Additionally, in Table 2, we compare the object shape parameters estimated by *HOPE* to the actual, physically measured ones, computed by averaging estimations for all multiframes of a given sequence. The standard deviation of these estimations is in the order of a few millimeters. It can be verified that the error in object shape estimation is satisfactory.

For *HOPE*, we also run a simple classification experiment. Although shape classification is not the focus of this work, it provides an indirect indication of the accuracy of the optimization process. For the sphere sequence (Fig. 6(a) and (b)), we ran *HOPE* assuming a cuboid, an ellipsoid and a cylinder. Table 3 shows the mean value and the standard deviation of the objective function of *HOPE* in all multiframes of the sequence. As it can be verified, the hypothesis of an ellipsoid better explains the observed scene. In fact, 98.67% of the multiframes were better explained by the ellipsoid, 1.33% by the cylinder and none by the cuboid.

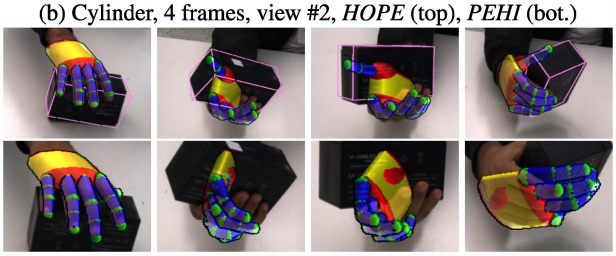
Finally, Fig.7, shows sample snapshots from the results obtained on a sequence of a hand performing fine manipulation of an elongated cuboid. Visual inspection confirms that the accuracy of *HOPE* is quite satisfactory, despite the complex and challenging hand-object interaction. Sample videos out of these experiments are provided as supplemental material to this submission and are also available online[2].

[2]http://www.youtube.com/watch?v=N3ffgj1bBGw



(a) Sphere, frame #103, 4 views, *HOPE* (top), *PEHI* (bot.)

(b) Cylinder, 4 frames, view #2, *HOPE* (top), *PEHI* (bot.)



(c) Box, 4 frames, view #1, *HOPE* (top), *PEHI* (bot.)

Figure 6. Sample frames from the results obtained by *HOPE* and *PEHI* in real-world experiments. For *HOPE* the projection of the estimated 3D object model is shown in pink color.

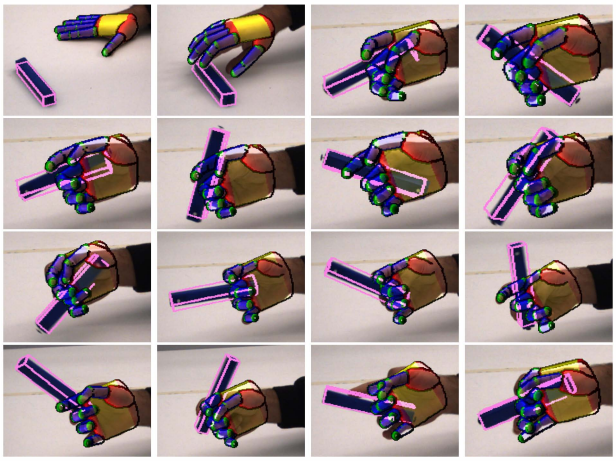


Figure 7. Snapshots from an experiment where a hand performs a complex manipulation of an elongated cuboid.

4. Discussion and conclusions

In a hand-object interaction scenario, the observation of hands provides information that is important to the understanding of the object's state and vice versa. In this paper,

we demonstrated that by considering jointly the hand and the object and by modeling occlusions and physical constraints it is possible to better understand aspects of both. More specifically, the optimization over the parameters of a joint hand-object 3D model results in full hand pose estimation that is performed more accurately compared to methods that consider the hand in isolation. On top of that, a parametric expression of the manipulated object is also computed. PSO is proved very competent in handling the complex, multidimensional and multimodal objective function of this problem. Results from extensive experiments on simulated data demonstrated the potential of the method against ground truth, but also comparatively to the results of a state-of-the-art hand pose estimation method that considers hands in isolation. Experiments in real world sequences exhibit that the proposed method performs well in challenging cases of complex hand articulation and hand-object interaction. Ongoing research investigates the potential of *HOPE* in supporting the interpretation of the semantics of human grasping and manipulation activities.

Acknowledgments

This work was partially supported by the IST-FP7-IP-215821 project GRASP. The contribution of Konstantinos Tzevanidis and Pashalis Padeleris, members of CVRL/FORTH, is gratefully acknowledged.

References

[1] I. Albrecht, J. Haber, and H. Seidel. Construction and animation of anatomically based human hand models. In *2003 ACM SIGGRAPH/Eurographics symposium on Computer Animation*. Eurographics Association, 2003. 3

[2] A. Argyros and M. Lourakis. Real-time tracking of multiple skin-colored objects with a possibly moving camera. In *ECCV*, 2004. 3

[3] V. Athitsos and S. Sclaroff. Estimating 3d hand pose from a cluttered image. In *CVPR*, volume 2, page 432, Los Alamitos, CA, USA, 2003. IEEE Computer Society. 2

[4] J. Canny. A computational approach to edge detection. *PAMI*, 8(6):679–698, 1986. 3

[5] M. Clerc and J. Kennedy. The particle swarm - explosion, stability, and convergence in a multidimensional complex space. *IEEE Transactions on Evolutionary Computation*, 6(1):58–73, 2002. 4

[6] M. de la Gorce, N. Paragios, and D. Fleet. Model-based hand tracking with texture, shading and self-occlusions. In *CVPR*, pages 1–8, 2008. 2

[7] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly. Vision-based hand pose estimation: A review. *CVIU*, 108(1-2):52–73, Oct. 2007. 2

[8] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *PAMI*, 31:1775–1789, 2009. 2

[9] A. Gupta, A. Mittal, and L. Davis. Constraint integration for efficient multiview pose estimation with self-occlusions. *PAMI*, 30(3):493 –506, Mar 2008. 2

[10] H. Hamer, K. Schindler, E. Koller-Meier, and L. V. Gool. Tracking a hand manipulating an object. In *ICCV*, Oct 2009. 2

[11] J. Kennedy, R. Eberhart, and Y. Shi. *Swarm Intelligence*. Morgan Kaufmann Publishers, 2001. 4

[12] H. Kjellstrom, J. Romero, D. Martinez, and D. Kragic. Simultaneous visual recognition of manipulation actions and manipulated objects. In *ECCV*, 2008. 2

[13] N. Kyriazis, I. Oikonomidis, and A. Argyros. A gpu-powered computational framework for efficient 3d model-based vision. Technical Report 420, FORTH, July 2011. 6

[14] Microsoft Corp. Redmond WA. Kinect for Xbox 360. 2

[15] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *CVIU*, 104(2-3):90–126, Dec 2006. 2

[16] I. Oikonomidis, N. Kyriazis, and A. Argyros. Markerless and efficient 26-dof hand pose recovery. In *ACCV*, 2010. 2, 3, 4, 5

[17] I. Oikonomidis, N. Kyriazis, and A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, Aug 2011. 2

[18] A. Oliva and A. Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520 – 527, 2007. 2

[19] J. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *ICCV*, page 612, Los Alamitos, CA, USA, 1995. IEEE Computer Society. 2

[20] J. Romero, H. Kjellstrom, and D. Kragic. Monocular real-time 3d articulated hand pose estimation. *IEEE-RAS Int'l Conf. on Humanoid Robots*, Dec 2009. 2, 3

[21] J. Romero, H. Kjellström, and D. Kragic. Hands in action: Real-time 3d reconstruction of hands in interaction with objects. In *ICRA*, 2010. 2

[22] R. Rosales, V. Athitsos, L. Sigal, and S. Sclaroff. 3d hand pose reconstruction using specialized mappings. In *ICCV*, 2001. 2

[23] L. Sigal and M. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *CVPR*, volume 2, pages 2041 – 2048, 2006. 2

[24] R. Smith. Open dynamics engine, http://www.ode.org/, 2006. 4

[25] B. Stenger, P. Mendonca, and R. Cipolla. Model-based 3d tracking of an articulated hand. *CVPR*, pages II–310–II–315, 2001. 2, 3, 4

[26] E. Sudderth, M. Mandel, W. Freeman, and A. Willsky. Visual hand tracking using nonparametric belief propagation. In *CVPR Wkshp on Generative Model-based Vision*, 2004. 2

[27] B. White and M. Shaw. Automatically tuning background subtraction parameters using particle swarm optimization. In *IEEE ICME*, 2007. 4

[28] Y. Wu and T. S. Huang. View-independent recognition of hand postures. In *CVPR*, pages 88–94, 2000. 2

[29] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, Jun 2010. 2