An Object-Dependent Hand Pose Prior from Sparse Training Data

Henning Hamer[1] Juergen Gall[1] Thibaut Weise[2] Luc Van Gool[1,3]

[1]Computer Vision Laboratory
ETH Zurich
{hhamer,gall,vangool}@vision.ee.ethz.ch

[2]LGG
EPF Lausanne
thibaut.weise@epfl.ch

[3]ESAT-PSI / IBBT
KU Leuven
luc.vangool@esat.kuleuven.be

Abstract

In this paper, we propose a prior for hand pose estimation that integrates the direct relation between a manipulating hand and a 3d object. This is of particular interest for a variety of applications since many tasks performed by humans require hand-object interaction. Inspired by the ability of humans to learn the handling of an object from a single example, our focus lies on very sparse training data. We express estimated hand poses in local object coordinates and extract for each individual hand segment, the relative position and orientation as well as contact points on the object. The prior is then modeled as a spatial distribution conditioned to the object. Given a new object of the same object class and new hand dimensions, we can transfer the prior by a procedure involving a geometric warp. In our experiments, we demonstrate that the prior may be used to improve the robustness of a 3d hand tracker and to synthesize a new hand grasping a new object. For this, we integrate the prior into a unified belief propagation framework for tracking and synthesis.

1. Introduction

Many tasks performed by humans require hand-object interaction. Be it grasping a cup, dialing a mobile phone or spraying an aerosol can, humans are accustomed to using their hands for the manipulation of objects and their eyes for observing such manipulation. By watching another person handle a single instance of an unknown object class, humans can easily imitate the observed hand poses to manipulate other instances of the same class. Although the strong correlation between the nature and shape of an object and the hand poses for its manipulation is obvious, only little work has been done so far to exploit this information for vision-based hand pose acquisition.

Hand tracking approaches either focus on freely moving hands for gesture recognition or regard a grasped object only as an occluder [22, 16, 34, 28, 1, 30, 29, 9]. For object handling, however, the degree of occlusion can be so large, that occlusion robustness alone is not sufficient. Due to missing observations, many spatial ambiguities for the phalanges occur that cannot be resolved without additional knowledge.

In the context of marker-less human motion capture, this issue has been addressed by introducing priors on motion patterns [27, 19, 31, 2] that are learned from a motion database. Although similar hand motion priors can be learned by acquiring a large motion database with data gloves, these priors will still not take the relation between object shape and hand pose into account.

Here a prior is proposed, which integrates such relation. We proceed in three steps:
1) A specific hand is captured in 3d while it manipulates a specific object of a certain object class. We map the captured poses, i.e. the 3d position and rotation of each hand segment (like a phalanx), into the local object coordinate system. Then, contact points on the object are detected. For illustration see Fig. 1. **2)** The knowledge coming from several observed manipulations - performed by different hands on different class members - forms the prior: a spatial distribution of the pose samples. **3)** The prior is generalized towards expected manipulations of new objects from that class, possibly by new hands, based on a geometric warping.

The adapted, object-specific prior can be used both for synthesizing grasps and improving tracking. Both tasks are embedded in the same belief propagation framework. A hand pose's probability can then be defined with respect to the prior, contact points, object intersection constraints, anatomical constraints of the hand, and data likelihood. A former version of this inference model, only considering anatomical constraints and the data likelihood, was presented in [9].

Inspired by the ability of humans to learn the interaction with an object from a single example, we focus on sparse training data, i.e. we can already build the prior by seeing only one instance of an object class being manipulated by one person.

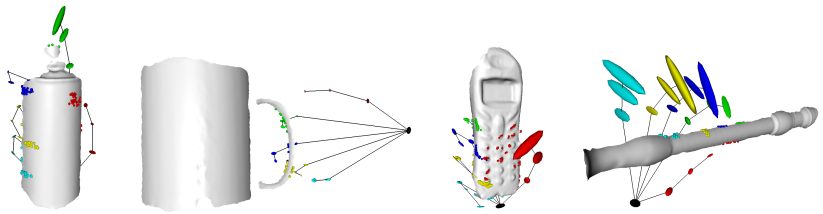


Figure 1. Captured object manipulation in a simplified illustration. An ellipse stands for a set of center points observed for some hand segment. Each finger has its own color. On the mesh, contact points of the individual hand segments are visualized as little dots.

2. Related Work

Vision-based hand pose acquisition is a challenging and active field of research. The majority of works have so far focused on various techniques to extract the hand pose without or with very limited additional scene information. For instance, tracking approaches that are based on local optimization [22], filtering [16, 28, 34], belief propagation [30, 9] or detections [1, 29] have been proposed. A detailed overview of state-of-the-art hand tracking is given in the survey [7].

The only priors that have been used so far in this context rely on the static or dynamic space of hand poses where a large dataset of hand shapes has been acquired by synthesizing hands or with data gloves [1, 29]. Since these works focus on gesture recognition, they consider only freely moving hands and cannot handle occlusions by an object. In the context of human pose estimation, priors on the human dynamics have been proposed that are learned from a motion capture database [27, 19, 31, 2]. While these priors do not capture the interaction with objects, basic constraints like contact with the ground plane have been used in [23, 32] for human pose estimation.

Object manipulation is targeted in the literature only in terms of action and object interpretation [13, 17] without exact pose estimation, so no detailed, high-dimensional model of the hand is provided. [13] for example focuses on the recognition of the general kind of manipulation and the manipulated object. Hand gestures are classified on a per-frame basis using 2d image features and learned sample sequences. [17] considers the relationship between objects, like an attachment or a contact between them, with the goal to explain the given scene. Hands are not treated separately and objects are recognized and tracked from 2d templates.

A taxonomy of human hand poses with regard to the grasping of objects has been provided in [5]. In [11], manipulative hand gestures are visually recognized using a state transition diagram that encapsulates task knowledge. The feature extraction is based on thresholding the hue value, so that the person has to wear special gloves, and gestures are simulated, without a real object being involved. [6] recognizes grasps referring to the grasp taxonomy defined in [5]. Real objects are handled, but do not impair the hand observation, because a data glove rather than visual input provides the hand pose.

Synthesis of grasps and hand motion has been addressed in the field of computer animation. Most similar to our approach are data driven approaches like [20, 14, 35]. In [20] a grasp controller has been proposed for a physically based simulation system. In order to obtain realistic behavior, the parameters of the controller are estimated from motion sequences captured with markers. A similar method is used in [14] where hand motion and contact forces are captured to estimate joint compliances. New interactions are synthesized by using these parameters for a physically based simulation. In [35] grasp synthesis is regarded as a 3d shape matching problem. A set of candidates is selected from a large grasp pose database by matching the contact points and surface normals of the hands and the object. Other approaches like [15] formulate the synthesis of hand manipulations as an optimization problem where an initial grasping pose and the motion of the object are given.

Grasps have also been studied in robotics [3]. Given a full 3d model and a grasp pose, for instance, the quality of the grasping can be evaluated based on pre-computed grasp primitives [18]. In [26], the 3d grasp position is estimated from two images where grasp locations are identified. For this, a 2d grasp point detector is trained on synthetic images. Other approaches are based on learning by demonstration and imitate human behavior. For instance in [10], a very small set of task relevant hand poses are selected and used to build a low dimensional hand model for grasp pose detection.

3. Prior

Statistical priors on hand poses are useful since they constrain the search space for tracking and allow for the prediction of hand poses when combined with additional constraints. The first property is important to overcome ambiguities due to missing data or occlusions and thus to improve tracking. The second property is important as well. In robotics, unseen instances of an object class need to be grasped. In computer graphics, the hand of an animated character should snap to a virtual object automatically. Since in both scenarios hand poses occurring during

object interaction are the most interesting ones, we aim to model a prior for the hand that depends on the object, i.e. we model the probability of a hand pose $\mathcal{P}$ conditional to an instance $\mathcal{O}$ of a known object class and a hand size $\mathcal{H}$: $p_{prior}\left(\mathcal{P}|\,\mathcal{O},\mathcal{H}\right)$. Before describing the prior in more detail, we briefly summarize our hand model $\mathcal{P}$.

3.1. Hand Model

In our hand model (Fig. 2(a)), each hand segment has its own 6-dimensional state space: three dimensions correspond to the position of the segment, three to its orientation (hand pose $\mathcal{P}$ has $16 \cdot 6 = 96$ DOFs). The state of a segment $\boldsymbol{x}_s \in \mathbb{R}^6$ is represented by a local coordinate system aligned with the segment. In addition, every phalanx is associated with a mesh approximating its skin for tracking. Each mesh is a composition of shape primitives like cylinders and spheres, with the exception of the more detailed thumb tip.

When modeling the hand by a set of individual segments, the likelihood of each segment s can be estimated independently with respect to simple local terms. Note that the connections between segments are used at a later stage to enforce anatomical correctness as explained in Sec. 4.

3.2. Prior Model

In consistency with the hand model, $p_{prior}\left(\mathcal{P}|\,\mathcal{O},\mathcal{H}\right)$ is defined by a product of local 6d hand segment distributions:

$$p_{prior}\left(\mathcal{P}|\,\mathcal{O},\mathcal{H}\right) = \prod_s p_{prior}\left(\boldsymbol{x}_s|\,\mathcal{O},\mathcal{H}\right). \qquad (1)$$

We learn the hand segment distributions from a finite set of hand segment samples $\boldsymbol{x}_s^i$, observed for instances $\mathcal{O}_k$ of the object class manipulated by the hands $\mathcal{H}_l$. For density estimation, we use a Parzen-Rosenblatt estimator with a 6d Gaussian kernel , defining $p_{prior}\left(\boldsymbol{x}_s|\,\mathcal{O},\mathcal{H}\right)$ by

$$\frac{1}{(2\,\pi\,\sigma^2)^{6/2}\,N}\sum_{i=1}^{N}\exp\left(-\frac{\left\|\boldsymbol{x}_s - f_{(\mathcal{O},\mathcal{H})}(\boldsymbol{x}_s^i)\right\|^2}{2\,\sigma^2}\right), \quad (2)$$

where N denotes the number of training samples $\boldsymbol{x}_s^i$, and σ is computed based on the max. nearest neighbor distance between all training samples. Since we estimate the probability of $\boldsymbol{x}_s$ conditional to $(\mathcal{O},\mathcal{H})$, we have to map the samples $\boldsymbol{x}_s^i$, observed conditional to $(\mathcal{O}_k,\mathcal{H}_l)$, into the coordinate system of object $\mathcal{O}$ and hand $\mathcal{H}$ by a warping function $f_{(\mathcal{O},\mathcal{H})}$. This mapping is explained in Sec. 5.

In analogy to [9], we model the overall probability of hand pose $\mathcal{P}$ as

$$p(\mathcal{P}) = \frac{1}{Z}\prod_{st}\psi_{st}(\boldsymbol{x}_s,\boldsymbol{x}_t)\prod_s\phi_s(\boldsymbol{x}_s), \qquad (3)$$

where the compatibility term $\psi_{st}(\boldsymbol{x}_s,\boldsymbol{x}_t)$ enforces anatomical constraints between adjacent hand segments, $\phi_s(\boldsymbol{x}_s)$ contains the data term with respect to the observation, and Z is a normalizing constant. A short introduction to the inference model with respect to hand tracking is provided in Sec. 4. The integration of the prior into $\phi_s(\boldsymbol{x}_s)$, and a unified framework for tracking and synthesis are the topic of Sec. 6.

4. Data Acquisition

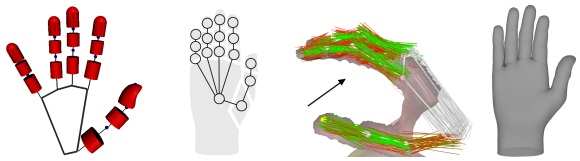
All our estimations are based on data retrieved by a structured light setup, delivering dense 2.5d range data and color information in real-time [33]. Using this system, we observe the manipulation of an object by a human hand and gather information regarding a) the fully articulated hand pose and b) the object's surface geometry, the object pose, and contact points on the object.

Hand Pose Our method exploits knowledge about the manipulating hand. For this, we use a hand tracker [9]. The tracker operates on a graphical model in which each hand segment is a node (Fig. 2(b)). First, depth information is evaluated locally for each hand segment to compute the data term $\phi_s(\boldsymbol{x}_s)$ in Eq. (3). Then, anatomical constraints between neighboring hand segments are introduced via the term $\psi_{st}(\boldsymbol{x}_s,\boldsymbol{x}_t)$. In each time step, samples are drawn locally around the hand segment states of the last time step (Fig. 2(c)), the observation model is evaluated, and belief propagation is performed[1] to find a globally optimal hand pose. For initialization, the hand pose is determined manually in the first frame.

Object occlusions make hand tracking a much harder task. Conceptually, the tracker is designed to handle this aggravated scenario. However, there are still situations in which the hand pose cannot be resolved correctly because the observation is corrupted to too large a degree. As a way out, we label the position of the finger tips in some key frames, making the training process semi-automatic. In those frames, finger tips are attracted by the labels instead of the local data. This said, the resulting prior eliminates any such manual intervention during testing.

Object Geometry and Pose As range scans of the object are captured continuously, we register these scans online and build up a coherent mesh of the already observed parts of the surface as demonstrated in [24]. Meshes obtained by this procedure are shown in Fig. 1. With the partial mesh of the object available, we determine in an offline process the object's 6d pose (translation and orientation) for each frame of a sequence containing the object and some kind of manipulation. This is done by fitting the mesh to the observation with ICP.

[1] using libDAI v0.2.2 (http://www.libdai.org)

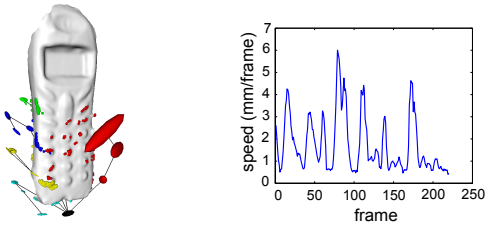


(a) Hand model (b) Graph (c) Local samples (d) 3d scan

Figure 2. (a) Hand model with a skeleton and ruled surfaces for the skin. (b) Graphical model for inference. (c) Depth data and hand segment samples. Color encodes relative observation likelihood: green is highest, red is lowest. The palm has uniform observation likelihood. An arrow indicates the viewing direction of the camera. (d) 3d scan of a hand used for hand pose visualization.

Given the hand pose and the object's geometry and pose, we find the closest vertex on the mesh for each hand segment. This vertex is saved as a contact point $\boldsymbol{c}_s^i$ of the respective segment if the distance is smaller than the segment's diameter. The partial mesh as well as the pose of the object are also required later on in our pipeline to compute hand-object intersection constraints.

Temporal Segmentation At this point, temporal segmentation is required to select the frames of interest. In action recognition, [21] identify action components with respect to velocity changes of a manipulating hand. We suggest to extend this concept to the velocity of individual hand segments. Consider Fig. 4 for a motivation.



(a) Manipulation (b) Speed of tip of thumb

Figure 3. (a) Captured hand segment centers and contact points. The thumb clearly is the most active finger and touches the dialing area at various positions. Considering the speed of the thumb tip plotted in (b), one could recognize the dialed phone number: when the thumb rests it is most likely to press on a digit.

In manipulation scenarios with less independent finger motion, we instead consider the global hand velocity in relation to the object's velocity, inspired by [25].

5. Warping

After having observed some hands manipulating some objects of a class, we can transfer the prior to another hand grasping another object of that class. We first adapt the acquired training examples $\boldsymbol{x}_s^i$ to the new hand size $\mathcal{H}$, and then warp them into the coordinate system of our newly observed object. The adapted prior (Eq. (1)) can be evaluated efficiently. We now describe these two steps of the warping process.

5.1. Hand Warping $\mathcal{H}$

Hand warping maps the samples to a new hand anatomy, i.e. adapts the length of the phalanges and the proportions of the palm. We preserve the position of the finger tips in space and elongate or shorten the finger segments, from farthest to closest to the palm, preserving joint angles. After this, the proportions of the palm (i.e. the relative positions of the attachment points of the five fingers) are set. Finger and palm adaption may create gaps between the fingers and the palm. We therefore apply the rigid motion to the palm that minimizes these gaps.

5.2. Object Warping $\mathcal{O}$

To map the prior from one object to a different one, we first warp the geometry of the corresponding meshes. For this, we use the method proposed in [36]. Corresponding points on the geometry have to be determined. We currently do this manually but a fully-automatic mechanism based on 3d features like spin images [12] could be realized. The output of mesh warping is an offset for each vertex of the original mesh that yields a point on the target mesh.

To apply the geometry warp to a hand segment sample, we need to map three points in 3d space (e.g. for a phalanx sample, the center, one end point, and one point fixing rotation around the roll-axis). This is necessary to fully define the warped 6d sample with respect to the coordinate system of $\mathcal{O}$. One approach to map a point in 3d space (outside the mesh) is to find the closest vertex of the original object's mesh and to choose the offset assigned to that vertex. While this might work in some cases, the accuracy was insufficient in our experiments. Instead, we use Radial Basis Functions (RBFs) [4] to extrapolate the warp field outside the mesh and move the hand segments with respect to this warp field.

6. Framework for Synthesis and Tracking

Since we have modeled our hand pose prior as a product of hand segment distributions (Eq. (1)), it is consistent with the hand tracking framework described in Sec. 4. This simplifies the integration of the prior to improve tracking. We use the same belief propagation framework not only for tracking but also for synthesis. Eq. (3) defines the probability of hand pose $\mathcal{P}$. In our case, $\phi(\boldsymbol{x}_s) = p(\boldsymbol{x}_s|I, \mathcal{O}, \mathcal{H})$, where I is a depth image, $\mathcal{O}$ an instance of an object with a pose and observed contact points. $\mathcal{H}$ defines the hand anatomy. According to Bayes' theorem,

$$p(\boldsymbol{x}_s|I, \mathcal{O}, \mathcal{H}) = \frac{p(I|\boldsymbol{x}_s, \mathcal{O}, \mathcal{H}) \cdot p(\boldsymbol{x}_s|\mathcal{O}, \mathcal{H})}{p(I|\mathcal{O}, \mathcal{H})}. \quad (4)$$

The denominator can be considered a normalization factor, as it does not contain $\boldsymbol{x}_s$. $p(\boldsymbol{x}_s|\mathcal{O},\mathcal{H})$ is defined by the object-dependent prior and augmented with two additional factors that enforce contact point attraction and intersection constraints:

$$p_{prior}(\boldsymbol{x}_s|\mathcal{O},\mathcal{H}) \cdot p_{contact}(\boldsymbol{x}_s|\mathcal{O},\mathcal{H}) \cdot p_{inter}(\boldsymbol{x}_s|\mathcal{O},\mathcal{H}).$$

Because the likelihood with respect to depth data is modeled as an exponential, we write

$$\phi(\boldsymbol{x}_s) = \frac{1}{Z} \exp\left(-\sum_{f=1}^{4} V_f\left(\boldsymbol{x}_s\right)\right), \tag{5}$$

For a detailed description of the likelihood $p(I|\boldsymbol{x}_s,\mathcal{O},\mathcal{H}) = exp(-V_1(x))$, we refer to [9]. The other terms are described next.

Hand Pose Prior The hand pose prior can be integrated in a straight-forward manner by taking the negative log probability of a sample with respect to the prior:

$$V_2\left(\boldsymbol{x}_s\right) = -\log\left(p_{prior}\left(\boldsymbol{x}_s|\,\mathcal{O},\mathcal{H}\right)\right). \tag{6}$$

Since the training samples are acquired from sequences of varying length, we weight the samples within the Parzen estimate (Eq. (2)) such that the sequences contribute equally to the prior.

Contact Point Attraction Allthough RBFs yield good results regarding warp extrapolation, small inaccuracies still occur when warping hand segment samples. Because of this, finger tips in contact with the original mesh do not always touch the mesh after warping. To yield stable grasps, we use contact points $\boldsymbol{c}_s^i$ observed on the original mesh. Since these contact points lie on the mesh they can be warped accurately without extrapolation. After warping, we proceed with 3d contact points as we did with 6d hand segment samples above and build a kernel estimate. The likelihood term $V_3\left(\boldsymbol{x}_s\right)$ of the distal phalanges with respect to the N_c contact points is then given by

$$-\log\left(\frac{1}{(2\,\pi\,\sigma_c^2)^{\frac{3}{2}}\,N_c}\sum_{i=1}^{N_c}\exp\left(-\frac{\left\|\boldsymbol{x}_s - f_{\mathcal{O}}(\boldsymbol{c}_s^i)\right\|^2}{2\,\sigma_c^2}\right)\right), \tag{7}$$

where $f_{\mathcal{O}}$ is the geometric warp. We again compute σ_c based on the max. nearest neighbor distance between training samples.

Intersection Constraints Intersection constraints concern hand segment samples that penetrate the mesh of the object after warping. We compute the smallest distance between the sample and the mesh. Because of the computational complexity, this is done for all samples in parallel on the GPU. Then, the degree of intersection of a sample with nearest vertex $v(\boldsymbol{x}_s)$ on the mesh is $d_{inter} = max(0, -(\|\boldsymbol{x}_s - v(\boldsymbol{x}_s)\| - d))$, where d corresponds to the diameter of the respective hand segment. We define

$$V_4\left(\boldsymbol{x}_s\right) = -\log\left(\frac{1}{Z}\exp\left(-\frac{{d_{inter}}^2}{{\sigma_{inter}}^2}\right)\right). \tag{8}$$

σ_{inter} is a user parameter, in our case set to 0.5. Note that there is no need to compute the normalizing constant Z, since it has no effect on belief propagation.

(a) Cup 1 (b) Cup 2 (c) Cup 3

Figure 4. Front side meshes of three different cups

Synthesis and Tracking In our experiments, we demonstrate that the prior can be used within the proposed framework to improve tracking and to synthesize hand poses for a given object. Tracking is performed as described in Sec. 4 but with the additional term of the pose prior. Grasp synthesis is realized within the same framework. For initialization, we consider all warped samples and choose one sample for each hand segment by belief propagation, maximizing the posterior of the global hand configuration. Warped samples do not necessarily result in anatomically valid hand configurations. Because of this, we then perform local sampling and belief computation several times. Within this belief propagation, anatomical constraints are enforced in the same way as during tracking.

7. Results

To evaluate our method, we have tracked the hand of seven different persons (one female) grasping, lifting, and putting down three different kinds of cups (Fig. 4)[2]. Based on the criterion regarding the temporal correlation of hand and object velocity (see Sec. 4 and Fig. 5), we selected those frames from the 21 sequences in which the hand firmly grasps the handle of the respective cup.

In the following we present two types of results to address the two application examples of our method: grasp synthesis and improved hand tracking.

7.1. Grasp Synthesis

In Sec. 6 we introduced three factors influencing the probability of a sample: consistency with the prior, con-

[2]data available at http://www.vision.ee.ethz.ch/~hhamer/cvpr10

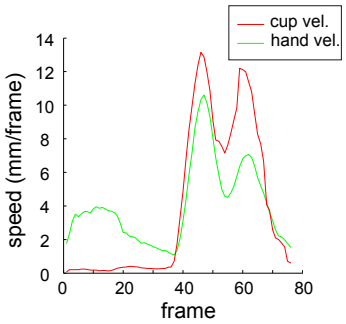
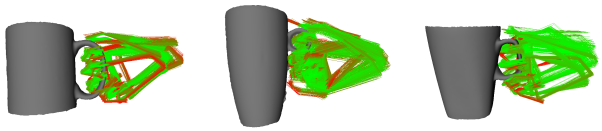


Figure 5. Speed of a cup and the manipulating hand. For temporal segmentation, we select the frames from about 40 to 80.

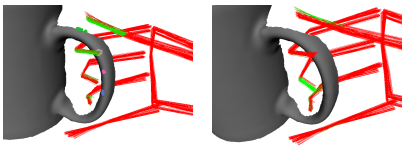
tact point attraction, and intersection constraints. The prior is visualized in Fig. 6 (a)-(c) for the different cup types. In these figures, the information of all seven tracked sequences is contained. With a 3GHz CPU and a GeForce 8800 Ultra, it takes $\approx$ 35 seconds to obtain each prior: 25 seconds to load in the database and to adapt hand anatomy, 10 seconds to transfer the prior to the target cup by warping.



(a) Cup 1 (b) Cup 2 (c) Cup 3

Figure 6. Prior for the three cup types. All seven test persons contribute to each prior. The variety of grasping is largest for cup 1 (two or three fingers in the handle), less for cup 3 (mostly two fingers in the handle) and least for cup 2 (anatomically, only one finger fits into the handle). The color of each samples encodes the probability with respect to the density defined by the prior itself (Eq. (6)). Red stands for a low and green for a high probability.

Fig. 7 gives examples with regard to the contact and intersection probabilities.



(a) Contact (b) Intersection

Figure 7. Probability terms favoring contact and avoiding intersection. Colors are normalized from red (low probability) to green (high probability), therefore all samples of hands segments without contact/intersection are drawn in red.

In Fig. 9 we show the process of synthesizing a grasp for a given cup and for a hand with a given anatomy (in this case the anatomy of an artificial hand: the 3d scan of a hand shown in Fig. 2(d), bound to a forward kinematic skeleton). The first image in each row shows the cup on which a grasp was actually observed. Contact points of the individual fingers are indicated as colored dots on the mesh. The rest of the rows illustrates the collected frames (adapted to the required hand anatomy), the transformed prior and the synthesized grasp (once rendered to visualize the 6d hand segment space and once using the artificial hand to allow for a better intuition of the results). Grasp synthesis based on a prior requires $\approx$ 30 seconds.

Fig. 10 shows the warp of a prior consisting of six of the seven sequences from cup 3 to cup 2. The figure demonstrates well the nature of our data-driven system: in all six sequences the test persons grasped the handle with two fingers. As a result, the hand synthesized for cup 3 exposes strong self-intersection of two fingers in the tiny handle (violating intersection constraints). However, the situation is resolved correctly by the system as soon as the seventh sequence (Fig. 10 (f)) is added to prior. Fig. 10 (g) and (h) show the final grasp.

7.2. Tracking

We now elaborate on results showing that a prior observed for one cup can improve hand tracking of a hand manipulating a different cup. Fig. 11 contains four frames of one of the 21 sequences. The handled object is cup 1 and the hand is the one of test person 2 (with a rather large hand). The prior we used in this experiment was observed on cup 3 and the manipulating hand (of test person 5) has average size. Without the labeling described in Sec. 4, tracking of the sequence fails due to strong ambiguity of the observation. The position of the distal phalanx of the middle finger significantly differs from the labeled ground truth and the data (see Fig. 12(a)). The red curve in Fig. 8 documents this. When we introduce the prior (Fig 12(b)) the middle finger remains in place (Fig. 12(c)) and we can track the sequence without any labels. The reduced error curve is also plotted in Fig. 8. Tracking with the prior and the hardware indicated above takes less then 10 seconds per frame.

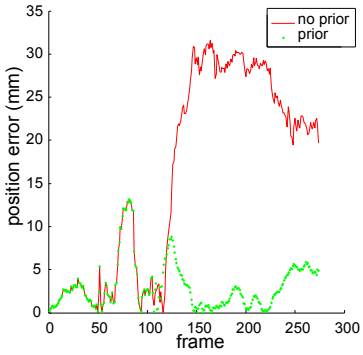


Figure 8. Tracking error of the distal phalanx of the middle finger. Without the prior obtained on the basis of a different cup and a different anatomy the error is significant. With the prior, the hand segment remains in place.

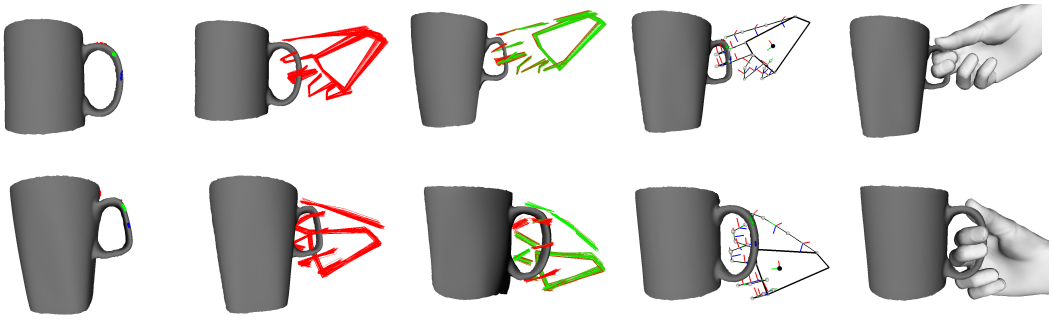
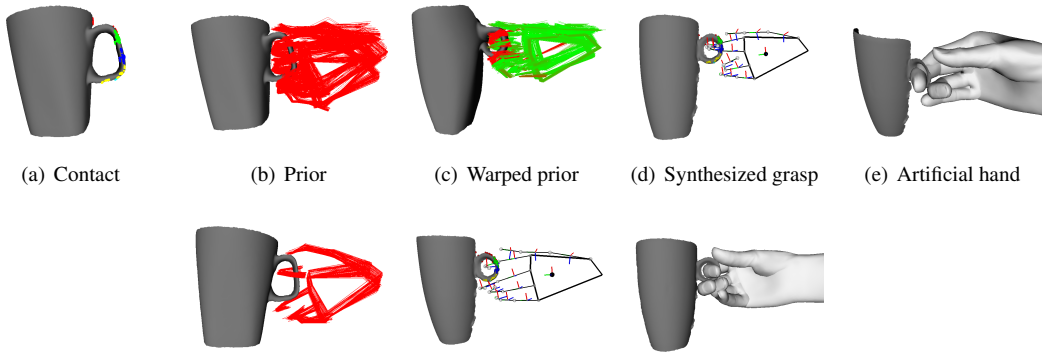


Figure 9. Synthesized grasps, derived on the basis of the observation of only one person grasping a different cup. (Row 1) Grasp of cup 3, using the observation of person 2 grasping cup 1. (Row 2) Grasp of cup 1, using the observation of person 1 grasping cup 3. (Col. 1) Originally observed contact points. (Col. 2) The derived prior. (Col. 3) The transfered prior. (Col. 4) The selected grasp (visualizing the 6d hand segment space). (Col. 5) The selected grasp (rendered using an artificial hand).

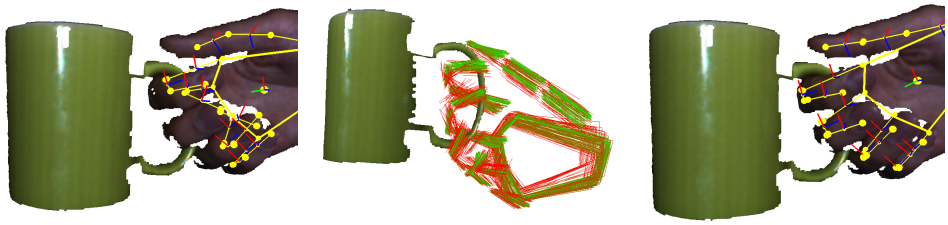


(a) Contact (b) Prior (c) Warped prior (d) Synthesized grasp (e) Artificial hand

(f) Additional samples (g) Final grasp (h) Result

Figure 10. (Row 1) Synthesized grasp of cup 2, based on six of the seven sequences observed for cup 3. All six test persons held the handle with two fingers. Consequently the resulting grasp strongly violates intersection constraints. As soon as we add sequence seven (f) (only one finger grasps the handle), the issue is resolved by our method (see (g) and (h)).

Figure 11. Four frames of the sequence showing person 2 grasping and lifting cup 1.



(a) Tracking failure (b) Applied prior (c) Successfully tracking

Figure 12. (a) While the cup is lifted, tracking fails due to strong ambiguates in the observation: the distal phalanx of the middle finger looses track. (b) The prior obtained from cup 3 and person 5. (c) The same frame as shown in (a), successfully tracked due to the prior, which stabilizes the tips within the handle.

8. Conclusion

The topic of this work is an object-specific hand-pose prior. The prior has a variety of applications and high relevance because of the great progress made in automated object class recognition.

We presented a unified framework for both hand tracking and grasp synthesis. While elaborating our pipeline we introduced ideas regarding temporal segmentation of action sequences, a method to warp object meshes and adapt hand poses, and a combination of factors realizing contact point attraction and mesh intersection avoidance.

In the result section we have successfully demonstrated our method on a dataset of 21 sequences, containing hands of seven different people manipulating three different kinds of cups. Firstly, we showed three grasps that were synthesized for unknown cups with only very sparse data. Secondly, we provided a quantitative evaluation showing that an object-dependent hand pose prior can improve the tracking of a hand manipulating an object. In the future we plan to apply our method to more object types and to combine our system with an automated object class recognition approach.

Acknowledgements The authors gratefully acknowledge support through the EC Integrated Project 3D-Coform. We also thank Raphael Hoever for sharing source code, and Joris Mooij for providing libDAI [8].

References

[1] V. Athitsos, S. Sclaroff. Estimating 3d hand pose from a cluttered image. *CVPR*, 2003.

[2] A. Baak, B. Rosenhahn, M. Müller, H.P. Seidel. Stabilizing motion tracking using retrieved motion priors. *ICCV*, 2009.

[3] A. Bicchi, V. Kumar. Robotic grasping and contact: A review. *ICRA*, 2000.

[4] M. Botsch, L. Kobbelt. Real-time shape editing using radial basis functions. *EUROGRAPHICS*, 2005.

[5] M. Cutkosky, P. Wright. Modeling manufacturing grips and correlations with the design of robotic hands. *ICRA*, 1986.

[6] S. Ekvall, D. Kragic. Grasp recognition for programming by demonstration tasks. *ICRA*, 2005.

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

[8] J. M. M. et al. libDAI 0.2.4: A free/open source C++ library for Discrete Approximate Inference, 2010.

[9] H. Hamer, K. Schindler, E. Koller-Meier, L. Van Gool. Tracking a hand manipulating an object. *ICCV*, 2009.

[10] M. Hueser, T. Baier, J. Zhang. Learning of demonstrated grasping skills by stereoscopic tracking of human head configuration. *ICRA*, 2006.

[11] K.-H. Jo, Y. Kuno, Y. Shirai. Manipulative hand gesture recognition using task knowledge for human computer interaction. *FG*, 1998.

[12] A. Johnson. *Spin-Images: A Representation for 3-D Surface Matching*. PhD thesis, CMU, Pittsburgh, PA, August 1997.

[13] H. Kjellström, J. Romero, D. M. Mercado, D. Kragic. Simultaneous visual recognition of manipulation actions and manipulated objects. *ECCV*, 2008.

[14] P. G. Kry, D. K. Pai. Interaction capture and synthesis. *TOG*, 25(3):872–880, 2006.

[15] C. K. Liu. Dextrous manipulation from a grasping pose. *SIGGRAPH*, 2009.

[16] J. Maccormick, M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. *ECCV*, 2000.

[17] R. Mann, A. Jepson, J. M. Siskind. The computational perception of scene dynamics. *CVIU*, 65(2):113–128, 1997.

[18] A. Miller, S. Knoop, H. Christensen, P. Allen. Automatic grasp planning using shape primitives. *ICRA*, 2003.

[19] K. Moon, V. Pavlovic. Impact of dynamics on subspace embedding and tracking of sequences. *CVPR*, 2006.

[20] N. S. Pollard, V. B. Zordan. Physically based grasping control from example. *SIGGRAPH*, 2005.

[21] C. Rao, A. Yilmaz, M. Shah. View-invariant representation and recognition of actions. *IJCV*, 50(2):203–226, 2002.

[22] J. M. Rehg, T. Kanade. Visual tracking of high dof articulated structures: An application to human hand tracking. *ECCV*, 1994.

[23] B. Rosenhahn, C. Schmaltz, T. Brox, J. Weickert, D. Cremers, H.-P. Seidel. Markerless motion capture of man-machine interaction. *CVPR*, 2008.

[24] S. Rusinkiewicz, O. Hall-Holt, M. Levoy. Real-time 3d model acquisition. *TOG*, 2002.

[25] Y. Sato, K. Bernardin, H. Kimura, K. Ikeuchi. Task analysis based on observing hands and objects by vision. *IROS*, 2002.

[26] A. Saxena, J. Driemeyer, A. Y. Ng. Robotic grasping of novel objects using vision. *IJRR*, 27(2):157–173, 2008.

[27] H. Sidenbladh, M. Black, L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. *ECCV*, 2002.

[28] B. Stenger, P. R. S. Mendonca, R. Cipolla. Model-based 3d tracking of an articulated hand. *CVPR*, 2001.

[29] B. Stenger, A. Thayananthan, P. Torr, R. Cipolla. Model-based hand tracking using a hierarchical bayesian filter. *PAMI*, 28(9):1372–1384, 2006.

[30] E. B. Sudderth, M. I. Mandel, W. T. Freeman, A. S. Willsky. Visual hand tracking using nonparametric belief propagation. *CVPR*, 2004.

[31] R. Urtasun, D. Fleet, P. Fua. 3d people tracking with gaussian process dynamical models. *CVPR*, 2006.

[32] M. Vondrak, L. Sigal, O. C. Jenkins. Physical simulation for probabilistic motion tracking. *CVPR*, 2008.

[33] T. Weise, B. Leibe, L. Van Gool. Fast 3d scanning with automatic motion compensation. *CVPR*, 2007.

[34] Y. Wu, J. Y. Lin, T. S. Huang. Capturing natural hand articulation. *ICCV*, 2001.

[35] L. Ying, J. Fu, N. Pollard. Data-driven grasp synthesis using shape matching and task-based pruning. *TVCG*, 13(4):732–747, 2007.

[36] L. Zhang, N. Snavely, B. Curless, S. M. Seitz. Spacetime faces: High-resolution capture for modeling and animation. *SIGGRAPH*, 2004.