2010 IEEE International Conference on Robotics and Automation
Anchorage Convention District
May 3-8, 2010, Anchorage, Alaska, USA

Hands in Action: Real-Time 3D Reconstruction of Hands in Interaction with Objects

Javier Romero Hedvig Kjellström Danica Kragic

Abstract— **This paper presents a method for vision based estimation of the pose of human hands in interaction with objects. Despite the fact that most robotics applications of human hand tracking involve grasping and manipulation of objects, the majority of methods in the literature assume a free hand, isolated from the surrounding environment. Our hand tracking method is non-parametric, performing a nearest neighbor search in a large database (100000 entries) of hand poses with and without grasped objects. The system operates in real time, it is robust to self occlusions, object occlusions and segmentation errors, and provides full hand pose reconstruction from markerless video. Temporal consistency in hand pose is taken into account, without explicitly tracking the hand in the high dimensional pose space.**

I. Introduction

Articulated tracking and reconstruction of human hands has received an increased interest within the fields of computer vision, graphics and robotics [1] and applications include learning from demonstration, rehabilitation, prosthesis development, human-computer interaction. Our goal is to equip robots with the capability of observing human hands in interaction with objects based solely on vision data, without markers.

Capturing hand articulation from video without markers is a challenging problem. A realistic articulated hand model has at least 28 degrees of freedom, making the state-space very large. The pose estimation suffers from self-similarity – fingers are hard to distinguish from each other – and a high degree of self-occlusion. Furthermore, hands move fast and non-linearly. Any method is thus computationally costly, making real-time implementation demanding. Although there are hand tracking systems developed for specific purposes such as sign recognition [1], full pose estimation remains an open problem, specially if real-time performance is required, as in virtually all robotics applications.

Hand pose estimation methods can largely be divided into two groups [1]: A) *model based tracking* and B) *single frame pose detection*. Methods of type A) usually employ generative articulated models [2], [3], [4]. Due to the high dimensionality of the human hand, they are facing challenges such as high computational complexity and singularities in the state space. They are thus generally unsuitable for robotics applications. Methods of type B) are usually non-parametric [5], [6]. They are computationally less demanding and more suited for a real-time system, but also more brittle and sensitive to image noise, since there is no averaging over time. In this paper we present a type B) non-parametric pose estimation method (Fig. 1), which takes temporal consistency into account. The probabilistic framework of this method is described in Section II. The method is faster and better at recovering from temporary errors than type A) model-based tracking methods. In an earlier paper [7] we also showed that the time continuity constraint makes the method more accurate and robust than other type B) single frame detection methods.

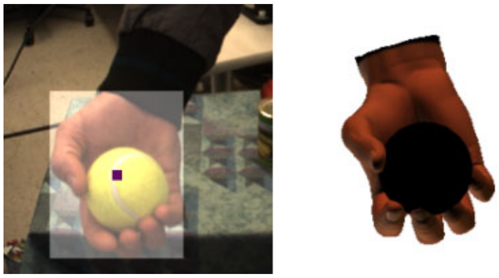


Fig. 1. Left) Original image and Right) Estimated pose.

The method maintains a large database of (synthetic) hand images. Each database instance is labeled with 31 parameters describing the hand articulation and orientation of the hand with respect to the camera. The 31D hand configuration of a new (real) image can then be found using an approximate nearest neighbor approach, taking previous configurations into account. Section II describes the composition of the database. The hand image representation is described in Section IV and the nearest neighbor-based mapping is described in Section V.

In the majority of applications, the human hands are frequently in contact with objects. Despite this, researchers have up to now almost exclusively focused on estimating the pose of hands in isolation from the surrounding scene. A recent notable exception is [8], who describe a type A) model-based tracker that allows for objects in the hand. Our method is also able to reconstruct hands both with and without grasped objects. Reconstruction of a hand grasping an object is in many ways a much more challenging task than reconstruction of a free hand, since the grasped object generally occludes large parts of the hand. The method of [8] allows for hand pose reconstruction *despite* the object occlusion.

On the other hand, knowledge about object shape gives important cues about the configuration of palm and fingers

This work is supported by EU through the project PACO-PLUS, IST-FP6-IP-027657, and GRASP, IST-FP7-IP-215821 and Swedish Foundation for Strategic Research.The authors are with the Computational Vision and Active Perception Lab, Centre for Autonomous Systems, CSC-KTH, Stockholm, Sweden. jrgn,hedvig,dani@kth.se

978-1-4244-5040-4/10/$26.00 ©2010 IEEE

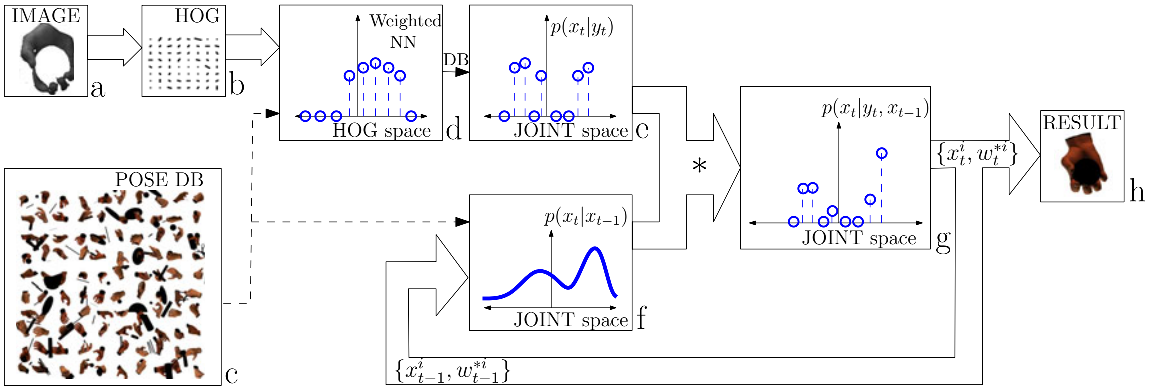


Fig. 2. The non-parametric temporal regression framework.

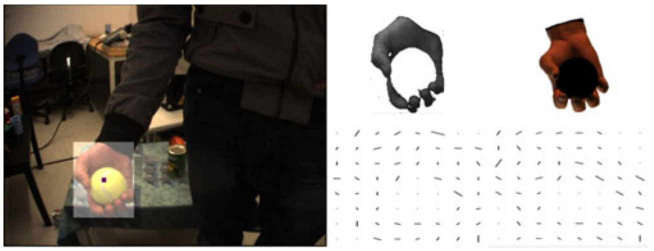
in contact with the object. Moreover, object shape and functionality give cues as to how this object is generally grasped. The relation between object shape and hand shape is however complex, and this information is hard to exploit in a type A) generative tracking model. In contrast to [8], our method is non-parametric, which means that complex object-hand shape dependencies can be implicitly represented by examples. Hand views in the database depicting grasping hands include occlusion from objects with a shape typical for this kind of grasp (Fig. 1). The occlusion affects the appearance of a hand view, so that hands with similar objects in them will appear similarly. Since the underlying assumption is that appearance similarity implies similarity in hand pose, the object shape *contributes* to the hand pose estimation in our method.

Thus, the main contribution of the paper is a robust non-parametric method for 3D hand reconstruction, operating in real-time, that also takes time continuity constraints into account. The method handles severe occlusions of the hand and also takes the object shape into account in 3D hand reconstruction. Experiments in Section VII also show that the method is robust to segmentation errors, a necessary requirement for the method to be applicable in a realistic setting.

II. Probabilistic Framework

The following notation is used throughout the paper. In a specific time instant t, let x_t be the articulated hand pose and y_t the observation. Here, x_t is a 28 dimensional vector of joint angles, and y_t is a 512D histogram of oriented gradients (HOG) [9], see Section IV. The space spanned by x is hereafter called JOINT space, while the space spanned by y is called HOG space. We assume that $p(x_t)$ is uniform over the JOINT space, and that the process is Markovian, i.e., x_t depends on the previous pose x_{t-1} only.

As shown in [7], the view y_t alone is not enough to non-ambiguously estimate the articulated hand pose x_t. Therefore, the pose x_{t-1} at the previous timestep is taken into account in the estimation. This corresponds to sequential estimation of $p(x_t|y_t, x_{t-1})$, the hand pose given the observation and the previous state. The temporal regression problem is decomposed as $p(x_t|y_t, x_{t-1}) \propto p(x_t|y_t)p(x_t|x_{t-1})$. As shown in Fig. 2, the method takes as input a monocular image and segments the hand based on skin color segmentation (a). A HOG y_t is then computed as described in Section IV (b).



(a) Original image (b) Segmented hand, HOG (c) NN in database, HOG

Fig. 3. Data representation.

The HOG y_t is compared to a large database of hand views (c), returning a weighted set of nearest neighbors $\{(y_t^i, x_t^i, w_t^i)\}$, as described in Section V (d). Each neighbor view y_t^i from the database has an associated joint angle configuration x_t^i, which, weighted by w_t^i, constitute a sampled approximation of $p(x_t|y_t)$ (e).

The temporal consistency constraint $p(x_t|x_{t-1})$ is a parametric function of x_t and x_{t-1}, as explained in Section VI (f). This term gives a higher probability to estimates where the hand has moved little over the last time step, thus giving priority to smooth motion estimates. The multiplication with $p(x_t|x_{t-1})$ is approximated by updating the database nearest neighbor weights to $w_t^{*i} \propto w_t^i p(x_t^i|x_{t-1})$ (g).

The expected hand pose value at time t is then estimated as $\hat{x}_t = E(x_t|x_{t-1}, y_t) \approx \arg\max_{x_t^i} w_t^{*i}$, i.e., the database pose with the highest weight (h).

III. Database Composition

The hand pose x_t could potentially be found by expressing $p(x_t|y_1, x_{t-1})$ parametrically, and finding the maxima of this function using an optimization algorithm. However, this optimization problem is high dimensional and non-convex. To alleviate the dimensionality problem, and constrain the search to commonly observed hand poses, we use a non-parametric approach: we discretize the state space by creating a large database of hand poses with synthetic images.

The composition of the database is motivated by our research aim: understanding human interaction with objects. Our database has more than 10^5 images, consisting of 5 different timesteps of 33 object grasping actions observed from 648 different viewpoints. The grasp types are selected according to the taxonomy presented in [10]. The graphics software Poser 7 is used to generate the synthetic hand views. The synthetic views in the database include basic object shapes that are usually involved in each kind of grasp (see Fig. 3c). The objects are considered background (although colored black for visibility in the figures) and the hand parts occluded by the object do not provide any features to the image observation y_t. This can be seen in Fig. 3c, bottom, where there is a "hole" in the middle of the HOG. As mentioned in the Introduction, the object shape contributes to the hand pose estimation in our method, since the hand pose depends on the shape of the object, which in turn affects the HOG y_t.

It can be argued that this method can only work if the object shape in the real action is the same as in the database. However, firstly, a particular kind of grasp is executed usually to similarly shaped objects and, secondly, the features used in our system (see Section IV) generalizes well over small variations in object shape. As described in Section II, $p(x_t|y_t, x_{t-1})$ is modeled non-parametrically using $\{(y_t^i, x_t^i)\}$, a set of database nearest neighbors to y_t in HOG space, weighted by their distance to y_t in HOG space and x_{t-1} in JOINT space. The weighting is formalized in Sections V and VI.

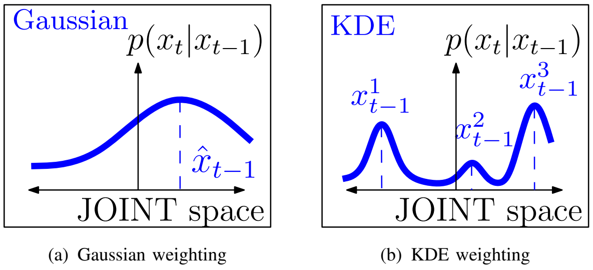
IV. Image Representation

The input to the method are monocular images of the type and quality shown in Figure 3a. In these images, the hand is segmented using skin color thresholding in HSV space [11] (Figure 3b, top). From the segmented hand image a histogram of oriented gradients (HOG) [9] is extracted (Figure 3b, bottom). This is a rich representation of shape, with certain robustness towards segmentation errors and small differences in spatial location and proportions of the segmented hand. The image is partitioned into cells and a histogram of gradient orientation is computed for each cell.

The size of the cells and the granularity of the histograms affect the generalization capabilities of the feature. With smaller cells and detailed histograms, the feature is richer but less capable of generalize over small differences. For our purposes, 8×8 cells and histograms with 8 bins provide good generalization with a sufficient level of details. The observation y_t equals the concatenation of the 8×8 histograms corresponding to each cell of the image. The dimensionality of y_t is thus $8 \times 8 \times 8 = 512$. A more detailed discussion on how different parameters of the HOG affect human detection can be found in [9].

V. Non-Parametric Mapping

The probability density function $p(x_t|y_t)$ is approximated by indexing into the database of hand poses using the image representation y_t, and retrieving the k nearest neighbors (kNN) in the space spanned by y.



(a) Gaussian weighting (b) KDE weighting

Fig. 4. Two different methods for modeling temporal consistency.

As an exact kNN search would put serious limitations on the size of the database, an approximate kNN search method, Locality Sensitive Hashing (LSH) [12] is employed. LSH is a method for efficient ϵ nearest neighbor (ϵNN) search. It is particularly suited for high dimensional data, since its online complexity does not depend explicitly on the set size or the dimensionality [12].

Each retrieved ϵNN y_t^i is given a weight $w_t^i = \mathcal{N}(y_t^i|y_t, \sigma_y)$, drawn from a 512D Gaussian density centered in y_t with standard deviation σ_y. This gives higher weight to database ϵNN that look similar to the observed hand.

In the database, each HOG y^j is associated with a pose x^j. The poses corresponding to the ϵNN $\{y_t^i\}$ can thus be retrieved. Together with the weights, they form the set $\{(x_t^i, w_t^i)\}$ which is a sampled non-parametric approximation of the density $p(x_t|y_t)$.

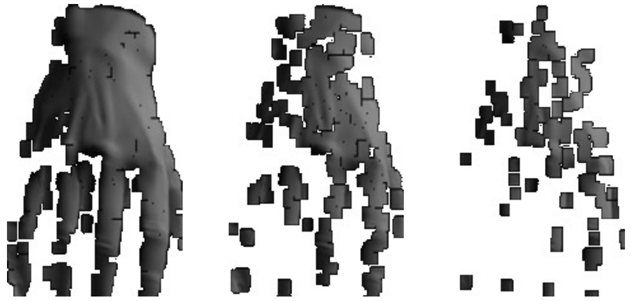
The pose vector x is composed of the rotation matrix of the wrist wrt the camera and the sines of the joint angles of the hand (which takes values between $[-\frac{\pi}{2}, \frac{\pi}{2}]$). Each component of x therefore lie in the domain $[-1, 1]$, which makes scaling unnecessary. The advantage of using a rotation matrix to represent the wrist rotation is that rotation matrices can be compared in a Euclidean fashion, as opposed to Euler angles and quaternions. Euclidean comparison of poses is used in the temporal consistency modeling (Section VI) and the experimental evaluation (Section VII-A).

VI. Temporal Consistency Modeling

As described in Section II, the temporal consistency constraint $p(x_t|x_{t-1})$ is modeled as a parametric function. It is used to reweight the sampled distribution $\{(x_t^i, w_t^i)\}$, approximating $p(x_t|y_t)$. We propose two ways to model the temporal consistency constraint, outlined in the two subsections below.

A. Single Hypothesis Gaussian Weighting

The simplest way of modeling temporal consistency is to assume that poses similar to the previous estimated pose $\hat{x}_{t-1}$ are more likely than poses that are very different from the previous one. Hence, $p(x_t|x_{t-1}) = \mathcal{N}(x_t|\hat{x}_{t-1}, \sigma_x)$, a 28D Gaussian density centered in $\hat{x}_{t-1}$ with standard deviation σ_x. This approach was used in [7] and is depicted in Figure 4a.



(a) $\alpha = 0.5\%$ (b) $\alpha = 3.3\%$ (c) $\alpha = 5\%$

Fig. 5. Artificial segmentation corruption α added to synthetic sequences.

B. Multiple Hypothesis Kernel Density Estimation Weighting

A drawback of the single hypothesis approach is that all the “second best" nearest neighbor hypotheses at $t-1$ are thrown away before temporal propagation. A logical improvement is to consider the full weighted set of hypotheses $\{(x^i_{t-1}, w^{*i}_{t-1})\}$ instead of the most likely hypothesis $\hat{x}_{t-1}$ in the estimation of $p(x_t|x_{t-1})$. This is illustrated in Figure 4b.

Following this idea, we use kernel density estimation (KDE) [13] over the weighted set of poses of the previous frame $\{(x^i_{t-1}, w^{*i}_{t-1})\}$ to estimate $p(x_t|x_{t-1})$. The system can then recover from an erroneous estimation of x_{t-1}.

As shown in the experiments in Section VII, KDE leads to a more robust sequential estimation than Gaussian weighting in many cases. Furthermore, even though KDE increases the computational load with a factor corresponding to the number of nearest neighbors $|\{x_{t-1}\}|$, the computational load of computing the temporal consistency weights is negligible compared to, e.g., the database ϵNN lookup. A drawback of KDE compared to Gaussian weighting is however the necessity of tuning more parameters, most importantly, the bandwidth of the kernels.

VII. Experiments

We first experimentally compare the two temporal consistency models detailed in Section VI, using synthetic sequences with hand pose ground truth. Then, the method is evaluated on real sequences featuring three different subjects and three object shapes. The sequences were captured at 10 frames/sec with a Point Grey Dragonfly camera with a resolution of 640×480 pixels. The method was implemented in C++ and runs at 10 frames/sec on one of the cores of a four core 2.66GHz Intel processor.

A. Comparison of Temporal Consistency Models

The single hypothesis and multiple hypothesis temporal consistency models are first compared in terms of pose reconstruction accuracy. This quantitative analysis of our method is done with synthetic sequences, where the hand pose ground truth is available. To make experimental conditions as realistic as possible, none of the hand poses or the objects in the synthetic sequences are present in the database. Moreover, the poses are corrupted with a variable amounts of segmentation noise (see Fig. 5), to simulate segmentation errors that occur with real sequences. The segmentation corruption is performed in the following way: The segmentation mask is first assigned as the full hand view (without noise). A fraction α of the pixels in the segmentation mask are set to zero. The error is then propagated through an erosion followed by dilation. In each frame t, the error of the estimated hand pose $\hat{x}_t$ relative to the ground truth pose x^{gt}_t is estimated as $||\hat{x}_t - x^{\text{gt}}_t||$, the Euclidean distance in the pose space explained in Section V. Figures 6 and 7 show the hand pose estimation of synthetic sequences 1 and 2 respectively, with segmentation corruption $\alpha = 0.5\%$.

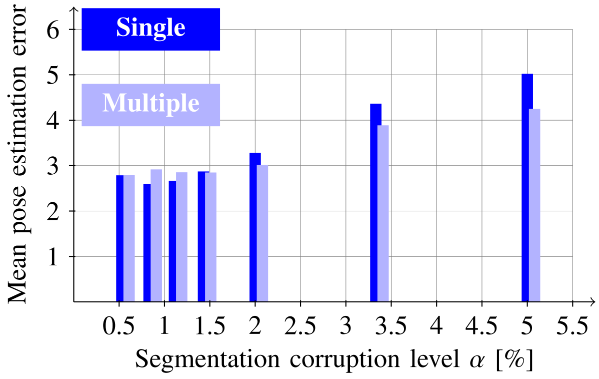


Fig. 8. Pose error with increasing segmentation corruption in sequence 1.

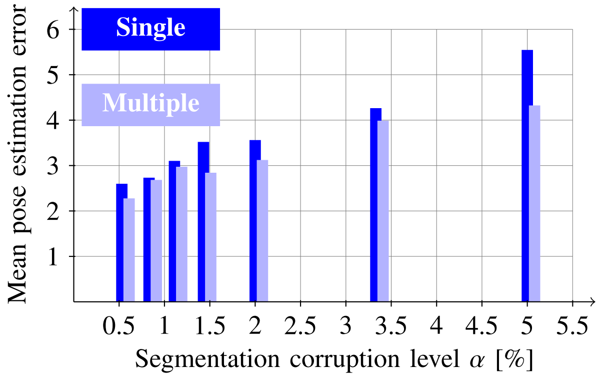


Fig. 9. Pose error with increasing segmentation corruption in sequence 2.

As shown in Fig. 8-9, the multiple hypothesis temporal consistency model almost consistently gives a better accuracy. The effect is more visible with higher segmentation corruption levels α. The reason for this is that the single-frame pose estimate $p(x_t|y_t)$ is more ambiguous for higher α, which means that there is a higher uncertainty about which sample x^i_t is the best pose estimate at time t. With higher α it is thus increasingly better to let all samples $\{(x^i_{t-1}, w^i_{t-1})\}$ influence the temporal model. It can als be seen that the pose estimation performance is largely unaffected by segmentation corruption levels up to $\alpha = 2\%$.

B. Real Sequences with Subjects Not in Database

To show the performance of the method on real data, it was evaluated with sequences of the first author and two uninstructed persons (one man and one woman) grasping

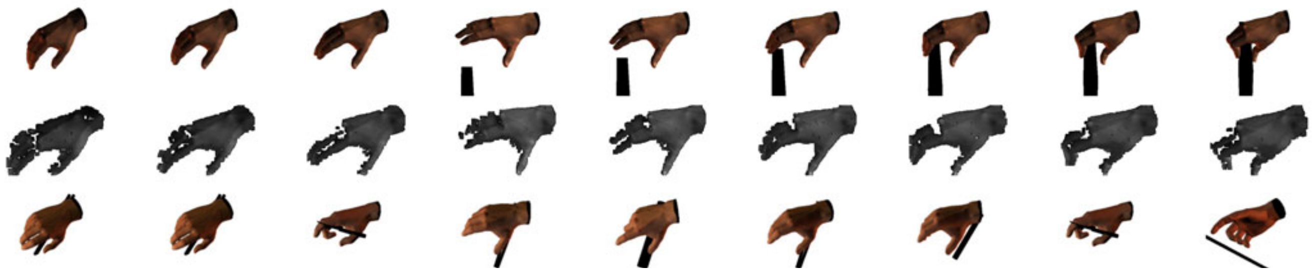


Fig. 6. Synthetic sequence 1. Top: original synthetic image. Middle: segmentation image with $\alpha = 0.5\%$. Bottom: estimated pose. *(The objects in the database are colored black for visibility here, but do not contribute to the HOGs.)* Video at www.csc.kth.se/~jrgn/VideosICRA2010/synthetic1.mp4

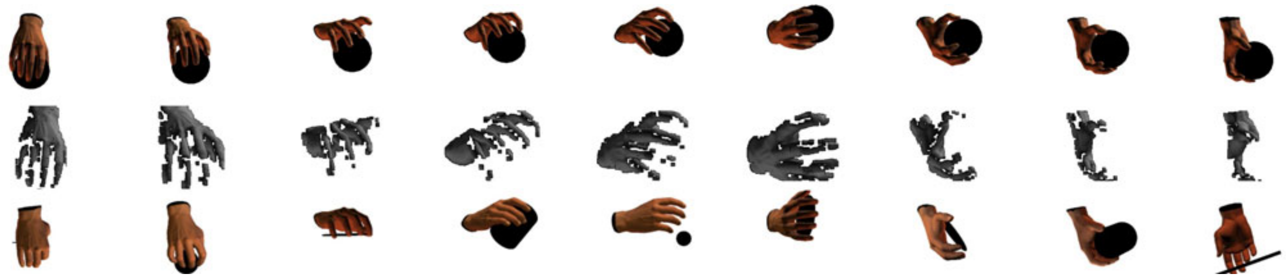


Fig. 7. Synthetic sequence 2. Top: original synthetic image. Middle: segmentation image with $\alpha = 0.5\%$. Bottom: estimated pose. *(The objects in the database are colored black for visibility here, but do not contribute to the HOGs.)* Video at www.csc.kth.se/~jrgn/VideosICRA2010/synthetic2.mp4

three different objects: A cup (with no equivalence in the database), a tennis ball (similar to a ball in the database), and a pair of pliers (with no equivalence in the database). The actions are not required to start from any specific pose. Naturally, the grasps in the sequences do not have exact correspondences in the database. Furthermore, the subjects' hands are of different sizes and shapes.

The multiple hypothesis temporal consistency modeling, shown above to be consistently better than the single hypothesis alternative, was used throughout the real image experiments. Fig. 10, 11, and 12, show the result of pose estimation for the three subjects respectively.

One conclusion that can be drawn is that the method is robust to individual variations in hand shape and proportions. The hand model used to generate the database view is designed to be male. However, the method is successful in recovering the poses of the considerably more slender female hand (Fig.”12), as well as of the hand with a larger proportion of the lower arm uncovered (Fig. 11); this affects skin segmentation, which in turn affects the HOG y_t used for database lookup.

The results also show that the method generalizes over grasps and objects that are not exactly represented in the database. It should be taken into account that two of the subjects have no previous experience with the method or the database, and thus can be expected to grasp the objects in a natural way. The cup and the ball are well represented by other objects present in the database. However, the pliers pose a slightly larger challenge for the method. There are two possible reasons for this. Firstly, the layout of the pliers, with two separated legs, makes the occlusion of the hand appear differently than any example in the database. Secondly, the functionality of the pliers makes the subjects grasp it differently than other rod-like structures in the database. Fig. 13 shows the pose estimation of a sequence where large parts of the hand is occluded by the grasped object showing the method is robust to large object occlusion.

The pose estimation in Fig. 14 points to an avenue for improvement of the method. In our current temporal continuity approaches we assume that the most probable current pose is similar to the most probable previous pose. With this we are making an implicit assumption of static hand pose. However, this assumption is frequently violated; fast hand motions like the one shown at the end of the sequence in Figure 14 are not uncommon. With the assumption of being static in the temporal consistency model, all poses x_t^i selected by the ϵNN sampling will be equally unlikely according to the temporal consistency model. Ambiguities in the HOG signature, e.g., between the front and back part of the hand, will then cause estimation errors as the one in the leftmost frame of Fig. 14. This issue can be addressed by including a dynamic model of pose over time.

VIII. Conclusions

A non-parametric method for 3D sequential pose estimation of hands in interaction with objects was presented. The contributions of this paper are the development of a method that not only handles severe occlusion from objects in the hand, but also takes the object shape into account in 3D hand reconstruction. In addition, the method is non-parametric and provides 3D hand reconstruction, operating in real-time, taking time continuity constraints into account.

Experiments showed that the method estimates hand pose in real time robustly against segmentation errors and large occlusion of the hand from objects. It was also shown that

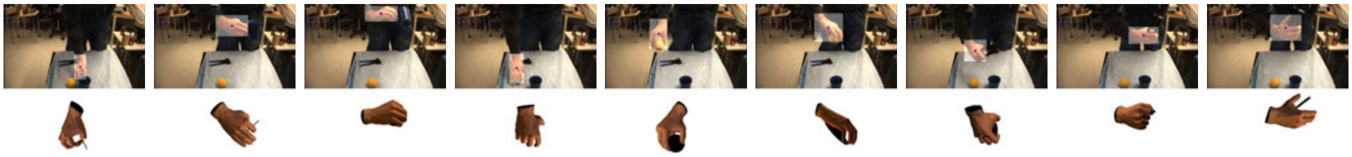


Fig. 10. Real sequence 1 (male subject 1). Top: image with skin segmentation window highlighted. Bottom: estimated pose. *(The objects in the database are colored black for visibility here, but do not contribute to the HOGs.)* Video at `www.csc.kth.se/~jrgn/VideosICRA2010/real1.mp4`

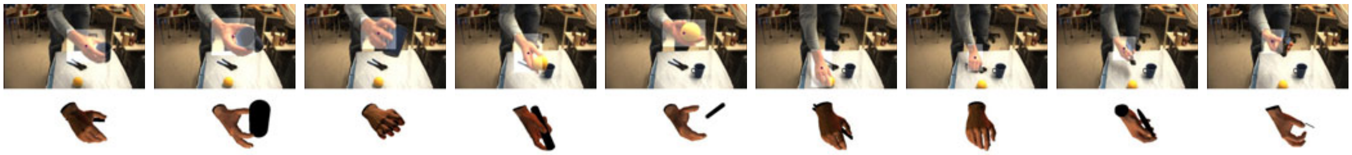


Fig. 11. Real sequence 2 (male subject 2). Top: image with skin segmentation window highlighted. Bottom: estimated pose. *(The objects in the database are colored black for visibility here, but do not contribute to the HOGs.)* Video at `www.csc.kth.se/~jrgn/VideosICRA2010/real2.mp4`

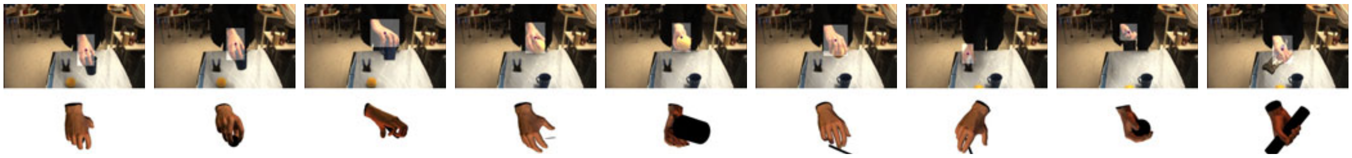


Fig. 12. Real sequence 3 (female subject 3). Top: image with skin segmentation window highlighted. Bottom: estimated pose. *(The objects in the database are colored black for visibility here, but do not contribute to the HOGs.)* Video at `www.csc.kth.se/~jrgn/VideosICRA2010/real3.mp4`

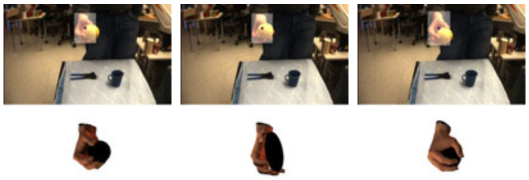


Fig. 13. Real sequence 4 (male subject 1) with large hand occlusion. Top: image with skin segmentation window highlighted. Bottom: estimated pose. Video at `www.csc.kth.se/~jrgn/VideosICRA2010/real4.mp4`

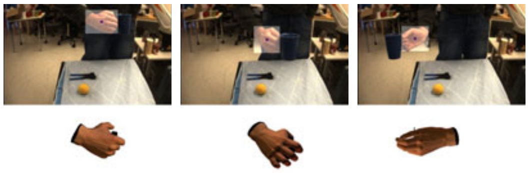


Fig. 14. Real sequence 5 (male subject 1) with fast non-linear motion. Top: image with skin segmentation window highlighted. Bottom: estimated pose. Video at `www.csc.kth.se/~jrgn/VideosICRA2010/real5.mp4`

the robustness to temporary estimation errors is improved by taking multiple hypotheses of previous hand pose into account.

Future work includes improving the motion model; currently, a static temporal model is implicitly assumed. This can be done in several ways, e.g., by learning low-dimensional models of hand motion from motion capture training data. Furthermore, we will enlarge the database to represent poses of differently shaped hands, grasping a wider range of objects under different illumination conditions. The approximate database lookup has a highly sub-linear time complexity, which allows for a significantly larger database with a moderate increase in computational load.

References

[1] A. Erol, G. N. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision-based hand pose estimation: A review," *Computer Vision and Image Understanding*, vol. 108, pp. 52–73, 2007.

[2] B. D. R. Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla, "Filtering using a tree-based estimator," in *IEEE International Conference on Computer Vision*, 2003.

[3] E. Sudderth, M. I. Mandel, W. T. Freeman, and A. S. Willsky, "Visual hand tracking using non-parametric belief propagation," in *IEEE Workshop on Generative Model Based Vision*, 2004.

[4] M. de la Gorce, N. Paragios, and D. J. Fleet, "Model-based hand tracking with texture, shading and self-occlusions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[5] V. Athitsos and S. Sclaroff, "Estimating 3D hand pose from a cluttered image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2003, pp. 432–439.

[6] H. Kjellström, J. Romero, and D. Kragić, "Visual recognition of grasps for human-to-robot mapping," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008.

[7] J. Romero, H. Kjellström, and D. Kragić, "Monocular real-time 3D articulated hand pose estimation," in *IEEE-RAS International Conference on Humanoid Robots*, 2009.

[8] H. Hamer, K. Schindler, E. Koller-Meier, and L. Van Gool, "Tracking a hand manipulating an object," in *IEEE International Conference on Computer Vision*, 2009.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. I: 886–893.

[10] T. Feix, R. Pawlik, H. Schmiedmayer, J. Romero, and D. Kragic, "A comprehensive grasp taxonomy," in *Robotics, Science and Systems Conference: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation*, June 2009.

[11] A. A. Argyros and M. I. A. Lourakis, "Real time tracking of multiple skin-colored objects with a possibly moving camera," in *European Conference on Computer Vision*, vol. 3, 2004, pp. 368–379.

[12] W. Dong, Z. Wang, M. Charikar, and K. Li, "Efficiently matching sets of features with random histograms," in *ACM Multimedia*, 2008.

[13] V. Morariu, B. Srinivasan, V. Raykar, R. Duraiswami, and L. Davis, "Automatic online tuning for fast gaussian summation," in *Neural Information Processing Systems*, 2008, pp. 1113–1120.