9th IEEE-RAS International Conference on Humanoid Robots
December 7-10, 2009 Paris, France

Monocular Real-Time 3D Articulated Hand Pose Estimation

Javier Romero Hedvig Kjellström Danica Kragic
Computational Vision and Active Perception Lab
Centre for Autonomous Systems
School of Computer Science and Communication
KTH, SE-100 44 Stockholm, Sweden
jrgn,hedvig,dani@kth.se

Abstract— **Markerless, vision based estimation of human hand pose over time is a prerequisite for a number of robotics applications, such as Learning by Demonstration (LbD), health monitoring, teleoperation, human-robot interaction. It has special interest in humanoid platforms, where the number of degrees of freedom makes conventional programming challenging. Our primary application is LbD in natural environments where the humanoid robot learns how to grasp and manipulate objects by observing a human performing a task. This paper presents a method for continuous vision based estimation of human hand pose. The method is non-parametric, performing a nearest neighbor search in a large database (100000 entries) of hand pose examples. The main contribution is a real time system, robust to partial occlusions and segmentation errors, that provides full hand pose recognition from markerless data. An additional contribution is the modeling of constraints based on temporal consistency in hand pose, without explicitly tracking the hand in the high dimensional pose space. The pose representation is rich enough to enable a descriptive human-to-robot mapping. Experiments show the pose estimation to be more robust and accurate than a non-parametric method without temporal constraints.**

I. Introduction

Vision based, markerless human hand tracking in natural environments with and without interaction with objects is an important building block for various human-machine interaction and robot learning tasks. An important aspect considered in our work is enabling robots to learn how to grasp and manipulate objects just by observing humans. Another aspect is monitoring of humans in everyday environments for designing hand prosthesis able of performing most common human grasps. However, capturing hand articulation is a challenging problem. Using the joint angle representation of hand pose requires 28-dimensional configuration space. In addition, self-occlusions of fingers introduce uncertainty for the occluded parts. Although there have been examples of systems that can track hands for very specific purposes such as sign recognition, full pose estimation remains an open problem, specially if real-time performance is required.

In robotic applications, an important aspect of task modeling is how different objects involved in the task should be grasped and manipulated. Humanoid robots are equipped with more and more dexterous humanoid hands, capable of perform human-like grasps. However, the control of these hands is far from trivial; therefore LbD is an attractive way of teaching the robot how to grasp [1]. While observing the human, the robot must estimate the human hand pose over time, and then map the hand pose to its own hands or grippers. In this paper we focus on visual estimation of human hand motion during object manipulation. While hand motion can be robustly extracted using 3D magnetic sensors or datagloves [2], the usability of a home service robot is compromised if the user is required to carry special markers during task instruction. The visual hand pose estimation is therefore required to be markerless.

(a) (b)

Fig. 1. a) ARMAR head, b) ARMAR head observing human grasp demonstration

Humanoid heads are constraint to have small baseline, lightweight stereo vision systems (see Figure I). This makes the stereo-matching problem difficult and sometimes inaccurate, specially for textureless surfaces as human hands. For this reason visual hand pose estimation based on monocular images can be an attractive field for humanoid robot research.

Markerless 3D reconstruction of hand pose based on a single image is an extremely difficult problem due to the large self-occlusion, high dimensionality and non-linear motion of the fingers. There are different ways of addressing these difficulties. Hand pose estimation method can largely be divided into two groups [3]: *model based tracking* and *single frame pose estimation*. Due to the high dimensionality of the human hand, articulated 3D model based trackers are facing challenges such as high computational complexity and singularities in the state space [4].Single frame pose estimation is usually more computationally efficient than model based tracking, but lacks the notion of temporal consistency, which is an important cue to hand pose [5], [6].

In earlier work [6], we presented a method for non-parametric estimation of grasp type and hand orientation

from a single monocular image. The method maintained a large database of (synthetic) hand images. Each database instance was labeled with the grasp type and the orientation of the hand with respect to the camera. The grasp type and orientation of a new (real) image could then be found using a nearest neighbor approach. For completeness, the hand image representation is described in Section III and the nearest neighbor-based mapping is described in Section IV.

In the current work, we have further developed the initial approach in two ways; I) by including temporal consistency in the distance measure used for database retrieval. This greatly enhances the robustness of the hand pose estimation, as it will be shown in Section VI; II) by extending the state space to a full joint angle representation, allowing a full 3D reconstruction of hand pose. This facilitates the learning of rich human-to-robot hand pose mapping. Development II) is the main contribution of this paper, described in more detail in Section IV and it is possible in part because of Development I), which is a secondary contribution, detailed in Section V.

Experiments in Section VI show that we can reconstruct the hand pose in real time and that our method is considerably robust to segmentation errors, a necessary requirement for the method to be applicable in a realistic setting. Additionally, it is shown that the temporal consistency constraint has a profound effect on the pose estimation accuracy and robustness.

II. Related Work

Analysis of human hand pose for the purpose of LbD [7] has been thoroughly investigated, almost exclusively with the help of markers and/or 3D sensors attached to the human hand [2]. However, we envision a LbD scenario were the teaching process can be initiated without calibration and where the robot-user interaction is as natural as possible. For this reason we want to reconstruct the hand posture in a visual markerless fashion.

The field of markerless visual hand pose estimation has been mainly devoted to hand gesture or sign language recognition [8]. A common approach is to estimate the hand pose from a single frame and use this pose as the input to a recognition module [5], [9], [10].The pose estimation is made easier by the fact that the range of poses can be constrained to the discrete set of specific gestures.

Methods for hand pose estimation that are not constrained to a limited set of poses can largely be classified into two groups [3]: I) model based tracking and II) single frame pose estimation. Methods of type I) usually employ generative articulated models [11], [4].Since the state space of a human hand is extremely high-dimensional, they are generally very computationally demanding, which currently makes this approach intractable for a robotics application. Methods of type II) are usually non-parametric [6].They are less computationally 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. The method presented here falls into the second approach. However, it takes temporal continuity into account and it can be used for on-line real-time reconstruction.

(a) HOG $\mathbf{x}$, JOINT $\mathbf{p}$ (b) HOG $\mathbf{x}_1$, JOINT $\mathbf{p}_1$ (c) HOG $\mathbf{x}_2$, JOINT $\mathbf{p}_2$

Fig. 2. Ambiguity in mapping from HOG space to JOINT space. Even though it is visually apparent that $\|\mathbf{p}-\mathbf{p}_2\| \ll \|\mathbf{p}-\mathbf{p}_1\|$ in JOINT space, database instance 1 will be regarded as the nearest neighbor as $\|\mathbf{x}-\mathbf{x}_1\| < \|\mathbf{x}-\mathbf{x}_2\|$. *Note that the object in the hand just contributes with occlusion of the hand in HOG extraction, as it is then colored uniformly with background color.*

For LbD purposes, it is relevant to investigate what hand pose information the robot needs in order to perform a successful human-to-robot mapping of the hand motion. In [12], [13] the control of a grasping hand was performed from a low dimensional space thanks to dimensionality reduction techniques.

III. Image Representation

The input to the method is a sequence $[\mathbf{I}_t], t = 1, \ldots, n$ of monocular images of the human hand. The same image representation was used in our previous work [6], where a more elaborate description can be found.

In each frame $\mathbf{I}_t$, the hand is segmented using skin color segmentation based on color thresholding in HSV space. The result is a segmented hand image $\mathbf{H}_t$. Due to a number of factors such as image noise, skin color in the background and non-skin colored areas on the hand (e.g. jewellery), the segmentation is more or less erroneous.

The shape information contained in $\mathbf{H}_t$ is represented with a Histogram of Oriented Gradients (HOG). This feature has been frequently used for representation of human and hand shape [14], [15].It has the advantage of being robust to small differences in spatial location and proportions of the depicted hand, while capturing the shape information effectively.

Gradient orientation $\mathbf{\Phi}_t \in [0, \pi)$ is computed from the segmented hand image $\mathbf{H}_t$ as $\mathbf{\Phi}_t = \arctan(\frac{\partial \mathbf{H}_t}{\partial y} / \frac{\partial \mathbf{H}_t}{\partial x})$.

From $\mathbf{\Phi}_t$, a pyramid with L levels of histograms with different spatial resolutions are created; on each level l, the gradient orientation image is divided into $2^{L-l} \times 2^{L-l}$ equal partitions. A histogram with B bins is computed from each partition.

The hand view at time t is represented by the HOG $\mathbf{x}_t$ which is the concatenation of all histograms at all levels in the pyramid. The length of $\mathbf{x}_t$ is thus $B \sum_{l=1}^{L} 2^{2(L-l)}$. Empirically, we obtained the best performance with a reasonable running time using $B = 8$ and $L = 3$. A discussion about how different parameters of the HOG affect human detection can be found in [16].

IV. Non-parametric Pose Reconstruction

In this section, we regard the problem of estimating a *single* pose $\mathbf{p}$ from a *single* HOG $\mathbf{x}$ omitting the time index.

The goal of the hand pose reconstruction process is to find the mapping $\hat{\mathbf{p}} = \mathcal{M}(\mathbf{x})$, where $\hat{\mathbf{p}}$ is the estimated 31D hand pose in terms of global orientation (lower arm yaw, pitch, roll) and joint angles (3 wrist joint angles, 5 joint angles per finger) , and $\mathbf{x}$ is the observed 168D HOG representation of the hand view, described in Section III.

The mapping function $\mathcal{M}$ can be expected to be highly non-linear in the HOG space, with large discontinuities. Following [6], $\mathcal{M}$ is therefore represented non-parametrically, i.e., as a database of example tuples $\{\langle \mathbf{x}_i, \mathbf{p}_i \rangle\}, i \in [1, N]$. Due to the high dimensionality of both the HOG space (168D) and the state space (hereafter denoted JOINT space, 31D), the database needs to be of a considerable size to cover all hand poses to be expected; in our current implementation, $N = 90000$. This has two implications for our mapping method, as outlined in the subsections below.

A. Generation of Database Examples

Generating a database of 10^5 examples from real images is intractable. Instead, we used the graphics software Poser 7 to generate synthetic views $\mathbf{H}_i^{\text{synth}}$ (see Figure 4) of different poses. The database was generated offline and it took around 5 days to render all the poses on a standard desktop computer. We are here motivated by the LbD application where we envision human to perform different types of grasps on objects in the environment. Therefore, the database examples are chosen as frames from short sequences of:

1) different grasp types, from
2) different view points, with
3) different grasped objects, and with
4) different illuminations.

The grasp types are selected according to the taxonomy developed in the GRASP project[1], which integrates the Cutkosky [17], Kamakura [18], and Kang [19] taxonomies. The whole database is also available at the same place. For each grasp type, a number of poses from whole grasp sequences (rest, approach and grasp) are included. Each pose is rendered with four different illuminations and from 386 different points of view uniformly distributed on a sphere. Standard objects are included to simulate typical occlusions.

From each example view $\mathbf{H}_i^{\text{synth}}$, the tuple $\langle \mathbf{x}_i, \mathbf{p}_i \rangle$ is extracted, where $\mathbf{x}_i$ is generated from $\mathbf{H}_i^{\text{synth}}$ as described in Section III, and $\mathbf{p}_i$ is the pose used to generate the view $\mathbf{H}_i^{\text{synth}}$ in Poser 7.

B. Approximate Nearest Neighbor Extraction

Given an observed HOG $\mathbf{x}$, the goal is to find an estimated pose $\hat{\mathbf{p}} = \mathcal{M}(\mathbf{x})$. With the non-parametric mapping approach, the mapping task $\hat{\mathbf{p}} = \mathcal{M}(\mathbf{x})$ is one of searching the database for examples $\langle \mathbf{x}_i, \mathbf{p}_i \rangle$ such that $\mathbf{x}_i \approx \mathbf{x}$. More formally, X_k, the set of k nearest neighbors to $\mathbf{x}$ in terms of Euclidean distance in HOG space, $d_i = \|\mathbf{x} - \mathbf{x}_i\|$ are retrieved.

As an exact kNN search would put serious limitations on the size of the database, an approximate kNN search method,

[1] www.grasp-project.eu.

Fig. 4. Synthetic sequence not contained in the database. *Note that the object in the hand just contributes with occlusion of the hand in HOG extraction, as it is then colored uniformly with background color.*

Locality Sensitive Hashing (LSH) [20] is employed. LSH is a method for efficient ϵ-nearest neighbor (ϵNN) search, i.e. the problem of finding a neighbor $\mathbf{x}_{\epsilon\text{NN}}$ for a query $\mathbf{x}$ such that

$$\|\mathbf{x} - \mathbf{x}_{\epsilon\text{NN}}\| \leq (1 + \epsilon)\|\mathbf{x} - \mathbf{x}_{\text{NN}}\| \quad (1)$$

where $\mathbf{x}_{\text{NN}}$ is the true nearest neighbor of $\mathbf{x}$.

The number of hyperplanes and number of tables used in the LSH search are learned from the database, as explained in [20]. In our current implementation, $K = 30$ and $T = 5000$.

The computational complexity of ϵNN retrieval with LSH [20] is $\mathcal{O}(DN^{\frac{1}{1+\epsilon}})$ which gives sublinear performance for any $\epsilon > 0$.

C. The Mapping $\mathcal{M}$ is Ambiguous

The database retrieval described above constitutes an approximation to the true mapping $\hat{\mathbf{p}} = \mathcal{M}(\mathbf{x})$, robust to singularities and discontinuities in the mapping function $\mathcal{M}$.

However, it can be shown empirically that $\mathcal{M}$ is inherently ambiguous (one-to-many); substantially different poses $\mathbf{p}$ can give rise to the similar HOGs $\mathbf{x}$ [14]. An example of this is shown in Figure 2.

Thus, the true pose $\mathbf{p}$ can not be fully estimated from a single HOG $\mathbf{x}$ (using any regression or mapping method); additional information is needed. In the next section, we describe how temporal continuity assumptions can be employed to disambiguate the mapping from HOG to hand pose.

V. Time Continuity Enforcement in JOINT Space

We now describe how temporal smoothness in hand motion can be exploited to disambiguate the mapping $\mathcal{M}$.

Consider a sequence of hand poses $[\mathbf{p}_t], t = 1, \ldots, n$, that have given rise to a sequence of views, represented as HOGs $[\mathbf{x}_t], t = 1, \ldots, n$. Since the mapping $\mathcal{M}$ is ambiguous, the k nearest neighbors to $\mathbf{x}_t$ in the database, i.e. the members of the set X_k, are all similar to $\mathbf{x}_t$ but not necessarily corresponding to hand poses similar to p_t. An important implication of this is that a sequence of hand poses $[\mathbf{p}_t], t = 1, \ldots, n$ does not necessarily give rise to a sequence of HOGs $[\mathbf{x}_t], t = 1, \ldots, n$ continuous in the HOG space. This is illustrated in the upper part of Figure 3, where we see that the red crossed arrow forcing continuity in HOG space points to the wrong pose.

This property of the data makes the problem of continuous hand pose recognition intrinsically different to other continuous NN problems found in the literature. For example, in [21] the "visible" feature displays time continuity, thus allowing the kNN answers from previous time steps to guide a new kNN query.

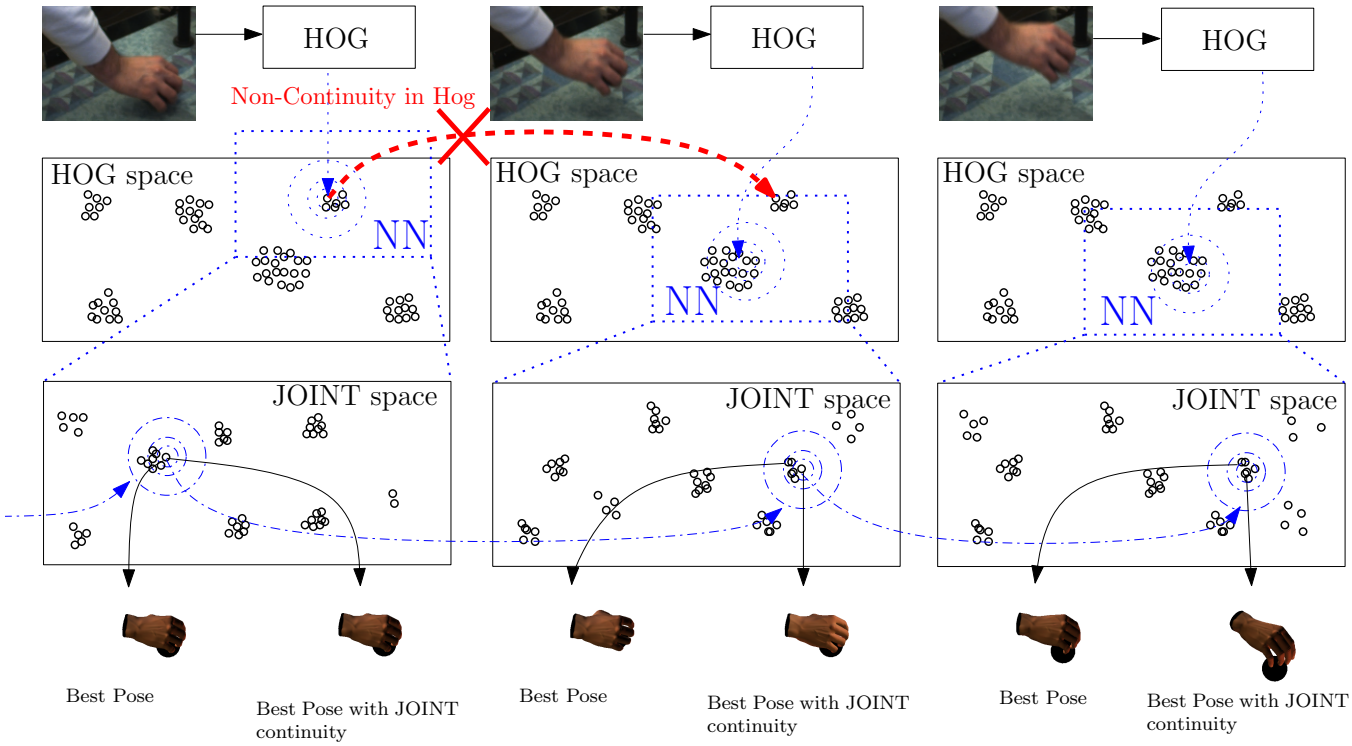
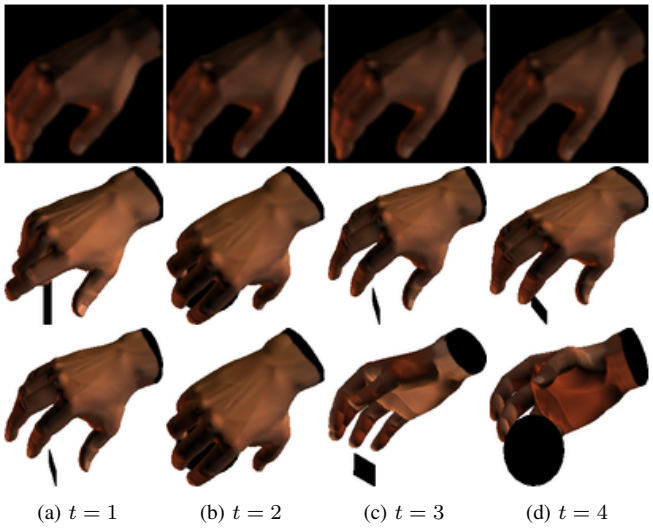


Fig. 3. Due to the underlying physics, a sequence of poses is continuous in the JOINT space, but not in HOG space.

However, due to the physics of the human body, the speed of the hand articulation change is limited. Thus, the sequence of hand poses $[\mathbf{p}_t], t = 1, \ldots, n$, i.e. the *hidden variables*, display a certain continuity in the JOINT space. This is illustrated in Figure 3.

The hand pose recognition for a certain frame t is therefore divided into two stages; I) retrieval of a set of k nearest neighbors X_k using single frame non-parametric mapping, as described in Section IV; II) weighting of the members of X_k according to their time continuity in the JOINT space.



(a) $t = 1$ (b) $t = 2$ (c) $t = 3$ (d) $t = 4$

Fig. 5. Recognition of hand pose with perfect segmentation. Row 1: query pose $\mathbf{p}_t$; Row 2: estimated pose $\hat{\mathbf{p}}_t$; Row 3: estimated pose $\hat{\mathbf{p}}_t^{\text{uniform}}$.

Let P_k be the set of poses corresponding to the kNN set X_k found in stage I). Moreover, let $\hat{\mathbf{p}}_{t-1}$ be the estimated pose in the previous time step. In stage II), the members $\mathbf{p}_j, j \in [1, k]$ of P_k are weighted as

$$\omega_j = e^{-\frac{\|\mathbf{p}_j - \hat{\mathbf{p}}_{t-1}\|}{2\sigma^2}} \quad . \tag{2}$$

where σ^2 is the variance of the distance from each entry pose $\mathbf{p}_j$ to the previous estimated pose p_{t-1}.

The pose estimate at time t is computed as the weighted mean of P_k:

$$\hat{\mathbf{p}}_t = (\sum_{j=1}^{k} \omega_j \mathbf{p}_j)/(\sum_{j=1}^{k} \omega_j) \ . \tag{3}$$

It should be noted that this is very similar in spirit to temporal filtering. The main difference is that a filtering approach can be regarded as *top-down*, making predictions about future poses according to some motion model, predicting how the observations of those prior poses should appear, and comparing the expected observations with the actual observations. Our approach can instead be regarded as *bottom-up*, making estimates directly from the observations, and then evaluating them in terms of the motion model.

In order to weight the poses p_j, p_{t-1} could be substituted by more complex predictions such as Kalman Filters or Particle Filters. However, the dynamics of the joints are not easy to model, so we preferred to keep the assumption about the dynamics as simple as possible as a first step. We leave the inclusion of a particle filter predictor for future work.

(a) $\alpha = 0\%$ (b) $\alpha = 1\%$ (c) $\alpha = 5\%$

Fig. 6. Synthesizing imperfect segmentation for synthetic images with 3 noise levels: fraction α pixels removed, followed by an opening-closing operation on the image.

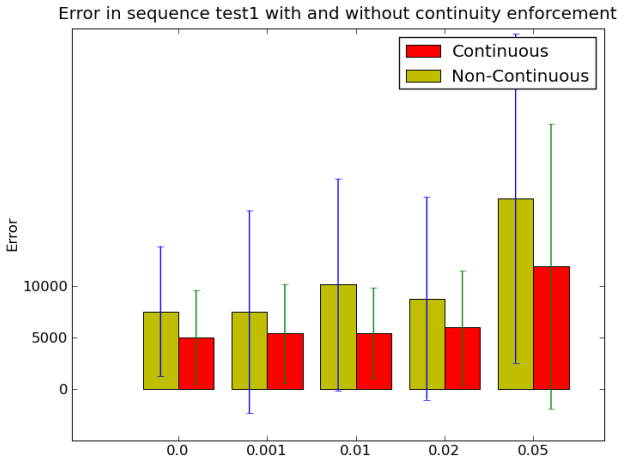


Fig. 7. Mean square error of 31D pose vector for continuous and non-continuous recognition

VI. Experiments

The experiments are designed to measure the effect of taking time continuity into account in the hand pose estimation as described in Equations (2), (3) as opposed to unweighted averaging

$$\hat{\mathbf{p}}_t^{\text{uniform}} = (\sum_{j=1}^{k} \mathbf{p}_j)/k \ . \tag{4}$$

Firstly, a quantitative analysis is made, using a synthetic sequence not included in the database. Secondly, the performance of the method is qualitatively evaluated on real images with hand poses not included in the database.

A. Quantitative Analysis

It is difficult to obtain ground truth poses $\mathbf{p}_t$ for a real image sequence; this would mean introducing markers, which would seriously affect the appearance of the hand. Therefore, a synthetic sequence is created, shown in Figure 4. The sequence depicts a typical approach-grasp action. Neither the rest position, the pose after the approach nor the final grasp pose are included in the database.

The quality of the estimated pose vector $\hat{\mathbf{p}}_t$ is measured in terms of Euclidean distance from the ground truth pose vector $\mathbf{p}_t$ in JOINT space: $E_t = \|\hat{\mathbf{p}}_t - \mathbf{p}_t\|$.

Fig. 8. General comparison. Row 1: query pose $\mathbf{p}_t$, not included in the database; Row 2: estimated pose $\hat{\mathbf{p}}_t$; Row 3: estimated pose $\hat{\mathbf{p}}_t^{\text{uniform}}$.

Figure 5 shows reconstructed poses $\hat{\mathbf{p}}_t$ compared to the baseline of $\hat{\mathbf{p}}_t^{\text{uniform}}$. The time continuity constraint is clearly effective: The estimates $\hat{\mathbf{p}}_t^{\text{uniform}}$ are much more incoherent over time than $\hat{\mathbf{p}}_t$. Figure 7, leftmost bar, shows that the mean error of sequence $[\hat{\mathbf{p}}_t], t = 1, \ldots, n$ is 50% lower than that of $[\hat{\mathbf{p}}_t^{\text{uniform}}], t = 1, \ldots, n$.

The comparison becomes more valid if we simulate realistic image noise conditions for this synthetic sequence. Noise is thus introduced in the segmentation of the image, in order to simulate imperfect segmentation in real sequences. This is done by removing a certain fraction of the pixels in the segmentation mask, followed by opening-closing morphological operations. Figure 6 shows how this operation affects the segmentation mask.

Figure 7 shows how the error (vertical axis) develops as the image segmentation noise level increases (horizontal axis). It is apparent that the estimation with pose continuity is much more robust to segmentation errors up to $\alpha = 2\%$. $\alpha = 5\%$ there is an abrupt increase in error for both methods, indicating that the segmentation (Figure 6c) then is too poor to yield descriptive HOGs.

B. Qualitative Analysis

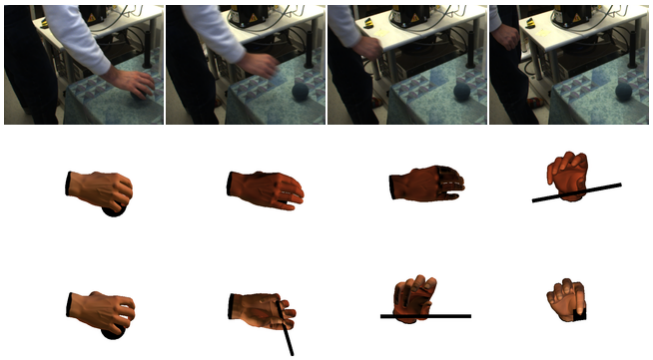
The algorithm was also evaluated with a real image sequence without known ground truth. The sequence contains grasps that do not correspond exactly to poses included in the database. Moreover, some grasps are performed with high velocity, yielding frames with substantial motion blur.

It should be noted that the experiments were performed with different people, only changing parameters of color skin segmentation. The system is quite robust to different hand shapes. The sequences were recorded with the ARMAR humanoid head (see Figure I). There is a decrease on performance when the hand occupies less than approximately 40x40 pixels.

Sample frames from the sequence are shown in Figure 8. The whole video with the results from the recognition system can be found at `http://www.csc.kth.se/~jrgn/handTracking264.mov`.

The main point of using continuity is to overcome ambiguity arising during a few frames, by taking into account past

Fig. 9. Segmentation error comparison. Column 1: query pose $\mathbf{p}_t$; Column 2: segmentation mask; Column 3: estimated pose $\hat{\mathbf{p}}_t$; Column 3: estimated pose $\hat{\mathbf{p}}_t^{\text{uniform}}$.



(a) $t = 1$ (b) $t = 3$ (c) $t = 5$ (d) $t = 7$

Fig. 10. Blurriness persistence. Row 1: query pose $\mathbf{p}_t$; Row 2: estimated pose $\hat{\mathbf{p}}_t$; Row 3: estimated pose $\hat{\mathbf{p}}_t^{\text{uniform}}$.

estimations. As expected, Figure 8 shows that the estimates $\hat{\mathbf{p}}_t^{\text{uniform}}$ are less robust to temporal ambiguities in the mapping $\mathcal{M}$. Enforcing continuity over time also improves the robustness towards motion blur and bad segmentation, as shown in Figures 10, 9. However, if the different problems (blurriness, poor segmentation) persist over more than 5–10 frames, the continuity enforcement does not contribute to the same extent.

Finally, we got some early results on a humanoid LbD scenario for grasping purposes [2].

VII. Conclusions

A non-parametric method for 3D hand pose estimation over time from a monocular video sequence was presented. Experiments showed that the system estimates the hand pose in real time robustly against segmentation errors. It was also shown that enforcing continuity in the hand pose space improves the quality of the hand pose estimation. Initial results showed that the system can be used in a LbD scenario for humanoid imitation.

Future work along these lines includes improving the motion model; currently, a static model is implicitly assumed. We can for example include angular velocities in the pose state space, thus encapsulating velocity information in the database examples. Furthermore, we will update the database to represent poses of differently shaped hands under different illumination conditions. We also plan to investigate methods for mapping the human hand pose to a lower dimensional space suitable for the robot hand that is going to actuate the grasp after LbD.

[2] http://www.csc.kth.se/~jrgn/GraspRecognitionDivx.avi

VIII. Acknowledgments

This project has been supported by the EU IST-FP6-IP PACO-PLUS, EU IST-FP7-IP GRASP and Swedish Foundation for Strategic Research through project CORS.

References

[1] S. Ekvall and D. Kragić, "Grasp recognition for programming by demonstration tasks," in *IEEE International Conference on Robotics and Automation*, 2005, pp. 748–753.

[2] L. Y. Chang, N. S. Pollard, T. M. Mitchell, and E. P. Xing, "Feature selection for grasp recognition from optical markers," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2007.

[3] A. Erol, G. N. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "A review on vision-based full DOF hand motion estimation," in *Vision for Human-Computer Interaction*, 2005, pp. III: 75–75.

[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] M. J. Matarić, "Sensory-motor primitives as a basis for imitation: Linking perception to action and biology to robotics," in *Imitation in Animals and Artifacts*, K. Dautenhahn and C. Nehaniv, Eds., 2000.

[8] V. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 677–695, 1997.

[9] T. Starner and A. Pentland, "Visual recognition of american sign language using hidden markov models," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 1995.

[10] 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.

[11] 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.

[12] M. T. Ciocarlie, S. T. Clanton, M. C. Spalding, and P. K. Allen, "Biomimetic grasp planning for cortical control of a robotic hand," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008, pp. 2271–2276.

[13] A. Tsoli and O. C. Jenkins, "Neighborhood denoising for learning high-dimensional grasping manifolds," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008, pp. 3680–3685.

[14] W. T. Freeman and M. Roth, "Orientational histograms for hand gesture recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 1995.

[15] G. Shakhnarovich, P. Viola, and T. Darrell, "Fast pose estimation with parameter sensitive hashing," in *IEEE International Conference on Computer Vision*, vol. 2, 2003, pp. 750–757.

[16] 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.

[17] M. Cutkosky, "On grasp choice, grasp models and the design of hands for manufacturing tasks," *IEEE Transactions on Robotics and Automation*, vol. 5, no. 3, pp. 269–279, 1989.

[18] N. Kamakura, M. Matsuo, H. Ishi, F. Mitsuboshi, and Y. Miura, "Patterns of static prehension in normal hands," *Am J Occup Ther*, vol. 7, no. 34, pp. 437–45, 1980.

[19] S. B. Kang and K. Ikeuchi, "Grasp recognition using the contact web," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1992, pp. 194–201.

[20] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," *Communications of the ACM*, vol. 51, 2008.

[21] R. Benetis, C. S. Jensen, G. Karciauskas, and S. Saltenis, "Nearest and reverse nearest neighbor queries for moving objects," *The VLDB Journal*, vol. 15, no. 3, pp. 229–250, 2006.