Image and Vision Computing 31 (2013) 555–564

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Non-parametric hand pose estimation with object context ☆,☆☆

Javier Romero [a,*], Hedvig Kjellström [b], Carl Henrik Ek [b], Danica Kragic [b]

[a] Perceiving Systems Department, Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany
[b] CVAP/CAS, KTH, SE-100 44 Stockholm, Sweden

ARTICLE INFO

Article history:
Received 5 January 2012
Received in revised form 26 January 2013
Accepted 11 April 2013

Keywords:
Articulated hand pose
Approximate nearest neighbor
Context

ABSTRACT

In the spirit of recent work on contextual recognition and estimation, we present a method for estimating the pose of human hands, employing information about the shape of the object in the hand. Despite the fact that most 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. Occlusion of the hand from grasped objects does in fact often pose a severe challenge to the estimation of hand pose. In the presented method, object occlusion is not only compensated for, it *contributes* to the pose estimation in a contextual fashion; this without an explicit model of object shape. Our hand tracking method is non-parametric, performing a nearest neighbor search in a large database (.. entries) of hand poses with and without grasped objects. The system that operates in real time, is robust to self occlusions, object occlusions and segmentation errors, and provides full hand pose reconstruction from monocular video. Temporal consistency in hand pose is taken into account, without explicitly tracking the hand in the high-dim pose space. Experiments show the non-parametric method to outperform other state of the art regression methods, while operating at a significantly lower computational cost than comparable model-based hand tracking methods.

1. Introduction

Human pose estimation is an important task for applications such as teleoperation and gaming, biometrics and prosthesis design, and human–robot interaction. However, accurate 3D reconstruction of human motion from images and video is a highly non-trivial problem, characterized by high-dimensional state spaces, fast and non-linear motion, and highly flexible model structures [2]. All this is applicable to hand reconstruction as well as full body reconstruction [1,3–6]. However, while a full body pose estimator encounters additional challenges from e.g. clothing, a hand pose estimator has to deal with other but equally demanding issues: similarity in appearance between the different parts of the hand (e.g. different fingers), and large self occlusion.

An important aspect of hand pose estimation is that humans are frequently interacting with objects. This is the case in the majority of the application areas mentioned above. The grasped object is often occluding a large part of the hand — for a plausible example, see Fig. 1, left.

Despite this, researchers have up to now almost exclusively focused on estimating the pose of hands in isolation from the surrounding scene, e.g. [7–11]. As illustrated in Fig. 1, top and middle, this will be inadequate if the observed hand interacts closely with objects during estimation.

Object–contextual hand pose estimation has been addressed in a generative manner in two recent works. In [12] the authors show that the hand pose can be reconstructed robustly despite the object occlusion. In [13], this is taken one step further, with explicit reconstruction of the object in 3D. By enforcing physical constraints on the hand pose from the object 3D surface and vice versa, the two pose estimation processes guide each other.

In contrast to [12,13], we take a discriminative approach to object–contextual hand pose estimation. The main contribution of this paper is a method for estimating human hand pose, employing contextual information about the shape of the object in the hand. Neither the hand nor the object is explicitly reconstructed; the hand and the object are instead modeled together, encoding the correlations between hand pose and object shape in a non-parametric fashion. In spirit of the recent methods for contextual recognition and estimation, e.g. [3,14,13,6], the object occlusion thereby helps in the hand pose reconstruction.

There are two reasons for exploring discriminative hand pose estimation with object context. Firstly, while generative estimation approaches commonly are more accurate, discriminative approaches are commonly more robust and computationally efficient; this is discussed further in Section 2. In, e.g., robotic and gaming applications, computational speed

☆ This paper has been recommended for acceptance by Ahmed Elgammal.
☆☆ This work was supported by the EU project TOMSY (ICT-FP7-270436) and by the Swedish Foundation for Strategic Research. An early version of the paper can be found in [1].
* Corresponding author. Tel.: +49 1774659896.
E-mail addresses: javier.romero@tuebingen.mpg.de (J. Romero), hedvig@kth.se (H. Kjellström), chek@csc.kth.se (C.H. Ek), dani@kth.se (D. Kragic).
URLs: http://ps.is.tuebingen.mpg.de/person/romero (J. Romero), http://www.csc.kth.se/~hedvig (H. Kjellström), http://www.csc.kth.se/~chek (C.H. Ek), http://www.csc.kth.se/~danik (D. Kragic).

0262-8856/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.imavis.2013.04.002

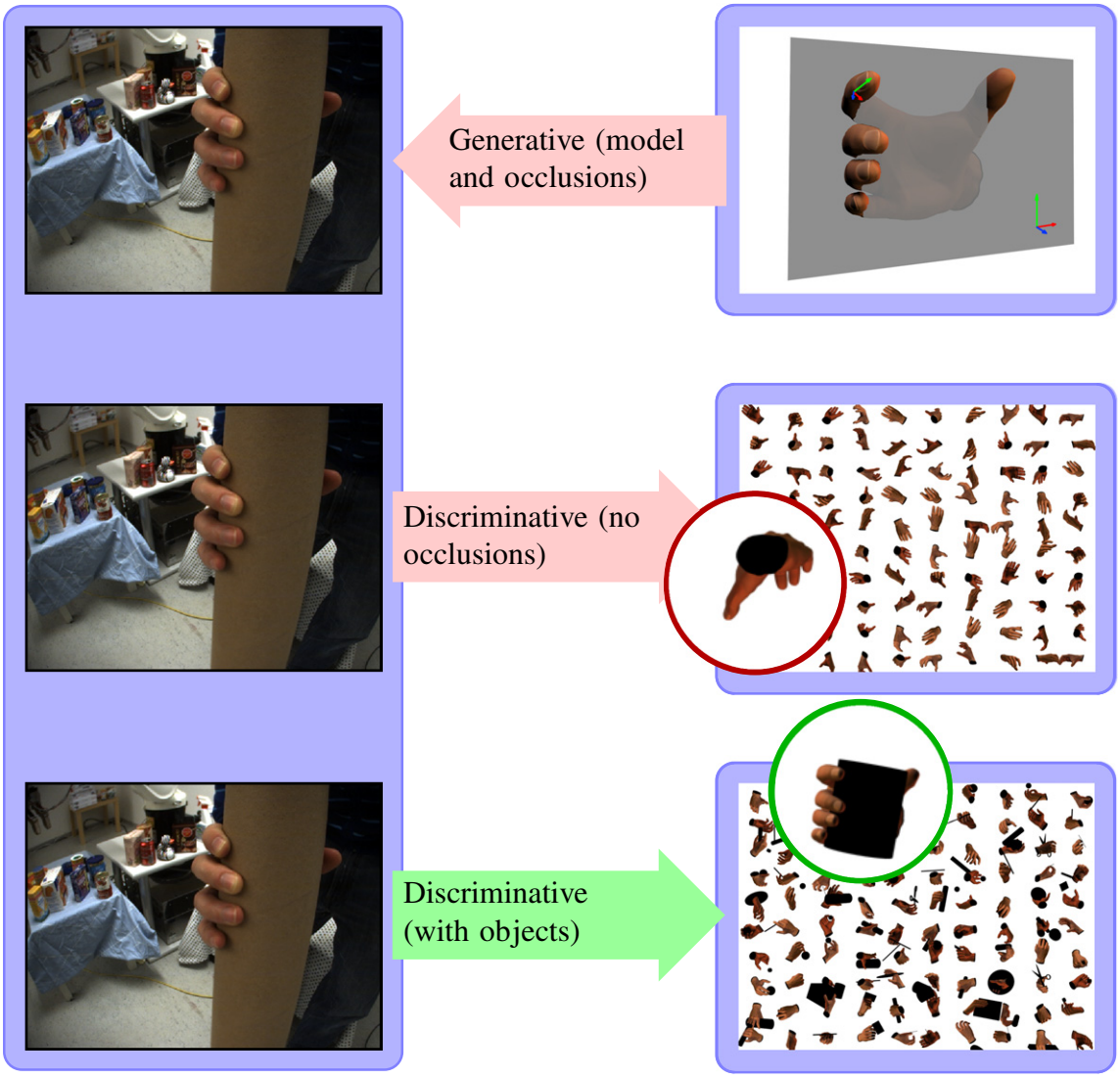


Fig. 1. Hand pose estimation is traditionally approached in two different manners, either with a generative model (top) or using a discriminative approach (middle). With a generative model, a model of the hand is maintained, and the image of the model is evaluated against the observed image. In a discriminative approach, the image generation process is not explicitly modeled; instead, a (parametric or non-parametric) mapping from image to pose is learned from training examples. If objects are not taken into regard in the modeling process, both these approaches have significant problems predicting in scenarios where large portions of the hand are occluded. In the generative case (top), there is too little image evidence to compute an informative likelihood. In the discriminative case (middle), the learned mapping can not take the object occlusion into regard, and will return an erroneous estimate. Our method (bottom) addresses this problem, by exploiting contextual information in the scene such as object–hand interaction. Due to this we can reliably predict pose in scenarios with significant occlusion. We would like to point out that our model is not limited to scenarios where an object is being manipulated but equally valid to estimate a free hand. Objects can also be taken into regard in a generative framework; see Section 2.

is critical, making discriminative approaches attractive. It is therefore valuable to investigate the possibility of estimating hand pose discriminatively in the context of objects.

Secondly, apart from the purely physical object constraints on the hand pose [13], there is also a functional correlation between object shapes and the manner in which they are grasped by a hand [15]. Thus, all physically possible ways of grasping an object are not equally likely to occur during natural object manipulation activities. Probability densities over hand pose conditioned on object shape can be encoded (in a non-parametric manner) in our discriminative method, while it is more difficult to encode this information in a generative model based method.

Fig. 1, bottom row illustrates our approach. In our non-parametric method, pose estimation essentially corresponds to matching an observed hand to a very large database (.. entries) of hand views. Each instance in the database describes the articulation and the orientation of the hand. The configuration of a new (real) image can then be found using an approximate nearest neighbor approach, taking previous configurations into account.

In our system, the database contains hands both with and without grasped objects. The database depicts grasping hands including occlusion from objects with a shape *typical for this kind of grasp*; this encodes functional correlations between object shape and the articulation of the grasping hand. The occlusion shape is strongly correlated to grasping type which further has a strong dependency with the hand articulation. Since the underlying assumption is that appearance similarity can be related to similarity in hand pose the object shape contributes to the hand pose estimation.

In many scenarios it is hard to differentiate between the palm and the dorsal ("back-hand") side of the hand. However, the object is much more likely to occlude the palm rather than the dorsal side of the hand. This gives insight on why object knowledge can be exploited in order to resolve the ambiguities typically associated with hand pose estimation. The rest of the paper is organized as follows: In Section 2

the relations to related work are discussed. The probabilistic estimation framework is then outlined in Section 3. The non-parametric hand model is described in Section 4, while Section 5 describes how inference is done over this model. Experiments in Section 6 show the non-parametric method to outperform other state of the art regression methods. We also show qualitative reconstruction results for a number of synthetic and real test sequences.

2. Related work

In this section we review related work on object–contextual non-parametric hand pose estimation. For a general review on human motion estimation we refer the reader to [2] and for hand pose estimation in specific to [16]. Further, we will discuss the main difference, both with respect to accuracy and performance, of generative and discriminative methods in the context of hand pose estimation.

2.1. Object–contextual hand pose estimation

As discussed in the introduction, hand pose estimation can be addressed in a generative or a discriminative manner. Object–contextual hand pose estimation has been addressed in a generative manner in two recent works. In [12] the authors show how the hand pose can be reconstructed robustly despite the object occlusion. The hand is observed using RGB-D image data. To achieve robustness to partial occlusion of the hand from objects, the hand is modeled as a Markov random field connecting segments corresponding to the different bones of the hand skeleton. In this way, the non-occluded segments can guide the pose estimation of the occluded ones.

In [13], this is taken one step further, with explicit tracking of the object in 3D. By enforcing physical constraints on the hand pose from the object 3D surface and vice versa, the two pose estimation processes guide each other. A multi-camera system is used to estimate both the pose of the hand and the object with frame rates between 0.5 and 2 Hz.

2.2. Generative and discriminative pose estimation

As outlined in the introduction inference of hand pose from images has either been done using generative or discriminative methods. In contrast to [12,13], we take a discriminative approach to object–contextual hand pose estimation. Over the next paragraphs we outline and discuss the main difference between generative model-based estimation methods and discriminative regression estimation methods to motivate our approach.

2.2.1. Accuracy

An important advantage of generative approaches is their (potential) accuracy, which is only limited by the precision of the hand model and the computational time available. In contrast, the accuracy of our discriminative non-parametric approach is fundamentally limited by the design of the database; it is not computationally tractable, using any approximation, to add enough new samples to the database in order to reach the accuracy of a generative tracker.

2.2.2. Initialization and error recovery

However, one disadvantage of generative models is their inherent local character. In most cases, the posterior distribution over the state space is highly multi-modal. The estimation procedure must therefore have a good prior state estimate. This can represent a problem in the initialization of the method. The tracking procedures in [12,13] were manually initialized.

Another inherent problem of locality with generative models is the recovery from errors; when the pose of a frame is wrongly estimated, subsequent frames will try to adapt such erroneous estimation to new frames. Since the temporal propagation model by nature is local, the method will then lose track.

Discriminative methods explore their full precomputed and discrete domain completely and independently every frame. This allows them to explore more efficiently broader sets of parameters compared to generative methods. In our system we encourage locality by using a temporal consistency model, see Section 5.2. However, since the likelihood in our model is sampled on a broad range of parameters, hypotheses from new parts of the pose space are continuously picked up, ensuring that the tracker can recover from errors easily.

The locality of model-based solutions can be specially problematic for hand pose estimation because hand movements in real sequences can be very fast (5 *m/s* translational and 300 *deg/s* rotational speed of the wrist [16]), breaking the locality assumption.

2.2.3. Computational efficiency

The joint estimation of hand and object pose in [13] presents another problem: computational load. The results shown with real sequences use eight cameras and the estimation time is 2 s per frame after speeding-up computations on the GPU. Decreasing the number of cameras (and therefore the quality) can speed-up the system up to 3 Hz. In [12] a running time of 6 s per frame is reported, although it is potentially parallelizable in a GPU.

In contrast, our discriminative method runs in real-time, implemented in C++ on a single CPU core. This allows other processes to run concurrently either in other CPUs or in the GPU, which is valuable for applications in robotics or gaming.

2.3. Non-parametric hand pose estimation

Other hand pose estimation systems have used databases of hand views in a non-parametric manner [7,8,11,17]. As discussed in the Introduction, none of the three previously mentioned systems mentioned how to handle or take advantage from occlusions, and the experiments showed hands moving freely without any object occlusion. The main difference between our system and previous approaches is that we exploit contextual information, such as objects to estimate the pose of the hand.

In [11], the application of a specially designed glove circumvents several problems associated with hand-pose estimation, making the problem as well as the approaches significantly different. An evolution of that system can be found in [17], where the authors track the hands without the need of gloves. However, they can only track a very limited range of hand poses and movements.

The system described in [7] performs the classification of human hand poses against a database of 26 basic shapes. This is adequate for their intended application, automatic sign language recognition. In contrast, our method aims to perform continuous hand pose estimation rather than isolated single-frame pose classification, which means that we can exploit temporal smoothness constraints to disambiguate the estimation.

The work from [8] can be regarded as the most similar to our work. However, like the two other approaches, they only take freely moving hands into regard.

3. Probabilistic framework

We begin by explaining the notation used throughout the paper. At a specific time instant t, let $\mathbf{x}_t$ be the articulated hand pose and $\mathbf{y}_t$ the corresponding image observation.

Given a specific image observation $\mathbf{y}_t$, we wish to recover the associated pose parameters $\mathbf{x}_t$ generating the visual evidence. Formally we will refer to the relationship between the pose and the image space as the *generative mapping* f,

$$\mathbf{y}_t = f(\mathbf{x}_t). \tag{1}$$

The task of pose estimation is to estimate the inverse of the generative mapping, either as a point estimate by modeling the inverse as a function, as in [18], or by a probabilistic method by estimating $p(\mathbf{x}_t|\mathbf{y}_t)$ which have the potential to handle a multi-modal estimate.

In the case of hand pose estimation, this is known to be a highly ill-conditioned problem, since the image features are ambiguous; the same image observation $\mathbf{y}$ might originate from a wide range of different poses $\mathbf{x}$, making the likelihood density multimodal [19]. In order to proceed, several different approaches have been suggested: generative models [20,12,13] which directly model f, approaches which rely on multiple views [9], or methods that exploit the temporal continuity in pose over time [20,21].

In this paper, our objective is a highly efficient method for situations where model-based generative approaches are inapplicable due to their computational complexity. Further, multiple views are not available in most applications.[1] We thus take the latter approach and exploit temporal continuity to disambiguate the pose. The pose space is assumed to be Markovian of order one, i.e., the pose $\mathbf{x}_t$ depends only on the pose at the previous time step $\mathbf{x}_{t-1}$. The estimation task thus reduces to find the pose $\mathbf{x}_t$ that maximizes $p(\mathbf{x}_t|\mathbf{y}_t, \mathbf{x}_{t-1})$ which decomposes as follows,

$$\arg\max_{\mathbf{x}_t} p(\mathbf{x}_t \mid \mathbf{y}_t, \mathbf{x}_{t-1}) = \arg\max_{\mathbf{x}_t} p(\mathbf{x}_t|\mathbf{y}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1}). \quad (2)$$

In this paper we take a non-parametric approach, with an implicit likelihood model represented by a large database of images and their corresponding poses, see Fig. 2. To perform inference, we use a truncated approach where we approximate the distributions in Eq. (2) using local models. As shown in Fig. 2, one time-step of inference is carried out as follows:

- Given an image observation $\mathbf{y}_t$, a set of weighted pose hypotheses $\mathbf{X}_t = \{\mathbf{x}_t^i, \mathbf{w}_t^i\}$ are drawn from the model as the nearest neighbors to the image observation in feature space. These constitute a sampled approximation of the observation likelihood $p(\mathbf{x}_t|\mathbf{y}_t)$. This is described in further detail in Section 5.1.
- From the weighted nearest neighbors of the previous time step, a function $g(\mathbf{x}_t)$ approximating the temporal model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is computed. This is described in further detail in Section 5.2.
- Weights $\mathbf{w}_t^{*i}$ are now computed as $\mathbf{w}_t^{*i} = g(\mathbf{x}_t^i) * \mathbf{w}_t^i$. The weights are normalized to sum to 1 for all samples in $\mathbf{X}_t$.
- The pose estimate is the most probable sample from the database given the observation and the previous estimates. With our weighted nearest neighbor approach, this is approximated by $\hat{\mathbf{x}}_t = \mathbf{x}_t^k$, where $k = argmax_i thbf w_t^{*i}$.

In the next section we describe how the proposed implicit database model is created and represented.

4. Non-parametric model representation

In order to obtain the non-parametric model, we need to acquire a training data set of poses and associated image appearances $(\mathbf{x},\mathbf{y})$ that can be assumed to "well" represent the problem, i.e., that includes poses that are expected to occur in a specific application domain. As our approach is non-parametric, there is no explicit parametrization of the image-to-pose mapping, as the relationship is implicitly parametrized by the database itself.

Generating such a database of natural images poses a formidable challenge, as it would need to capture the variations in pose and image appearance at a sufficient resolution in order to make accurate pose estimation possible. However, with recent advances in Computer Graphics we can use a rendering software such as Poser, which is capable of generating high-quality images of hands efficiently. The idea of acquiring large sets of training data using this approach is not new and has proved to be very successful for pose estimation [22,4].

The composition of the database used in this paper is motivated by our research aim: understanding human interaction with objects [23,24,14]. We select 33 different grasping actions according to the taxonomy presented in [15] (see one example in Fig. 3, left). Further, each action is applied to a set of basic object shapes on which the grasp would naturally be applied. Each action is then discretized into 5 different time-steps. In order to make our approach view-independent we generate samples of each instance from 648 different view-points uniformly located on the view-sphere. This results in a database of over 100 000 instances, which we assume samples the problem domain well.

4.1. Data collection

Images are extremely high-dimensional objects, making it infeasible both in terms of storage and modeling to use the original pixel representation. In this paper we therefore apply a two stage feature extraction approach with the aim to remove variance not related to pose from the image. In the first stage the hand is segmented from the image using skin color thresholding [25]; this also removes the object being grasped and the parts of the hand occluded by the object. This stage assumes that the object is not skin-colored. The system should be robust to objects with small skin-colored patches, since the effect should be similar to segmentation noise as explored in Section 6.1. Uniformly skin-colored objects are not considered in our approach. This assumption can be relaxed in different ways that compromise certain features of our system and go beyond the scope of this paper, for example model-based object tracking (but the system would lose the ability to handle unknown objects) or movement-based object tracking (under the assumption of the person and object being the only moving parts of the scene). Having extracted the hand from the image, the dimensionality is further reduced by representing the image as the response to an image feature.

A large amount of work within Computer Vision has been focused on developing different image features [26–28]. An ideal image feature should be robust to segmentation errors, sensitive to non-textured regions and fast to compute. We compare the performance of Histogram of Oriented Gradients (HOG) [29] features and features based on distance transform [30] for different parameter settings. For a number of different feature options, the following experiment is performed: The feature is computed for every database entry. The entries are removed from the database one at a time, and the 50 nearest neighbors (NN) extracted from the database. The mean is taken of the Euclidean distance in pose space between all query entries and their found nearest neighbor number 1, 2, …, 50. This distance is the same as the error of a non-parametric pose estimation — a dense database and a good feature would give small distances, while a sparse database and a non-informative feature would give large distances. Fig. 4 shows the cumulative mean pose error of nearest neighbor number 1–50, for 9 different feature alternatives.

Based on the result shown in Fig. 4, an $8 \times 8 \times 8$ HOG feature is selected, resulting in a 512 dimensional image representation, see Fig. 3, right.

Our motivation is to exploit contextual information of the grasped object when estimating the hand pose; the object contains a significant amount of information about the pose (and vice versa). In a learning based framework, which assumes having a training data set which describes the problem domain well, the natural inclination is that the model would be limited to handle objects which are included in the database. Such a model would have to be of a size that would render it infeasible to use. However, in our model the object is removed (assuming it is not uniformly skin-colored). This means the occluding shape of the object affects the representation while the internal edges of the object do not, see Fig. 3. This representation can robustly

[1] It should be noted that it is straight-forward in the present approach to employ image evidence from several camera views, or alternatively from RGB-D imagery. This is also discussed in the Conclusions.

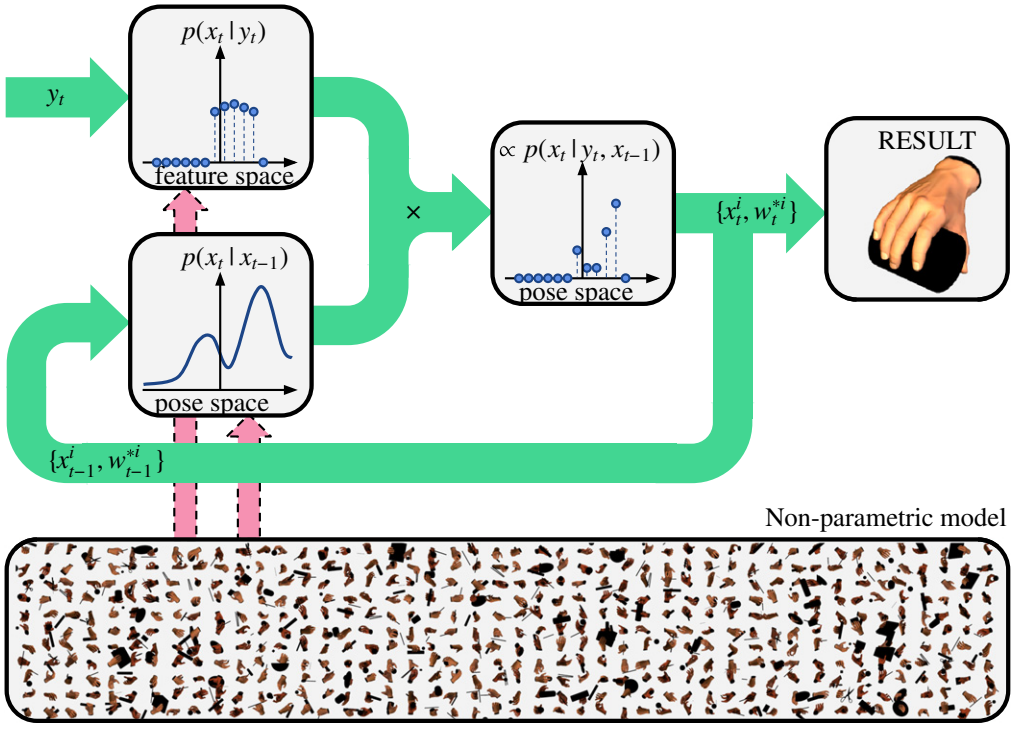


Fig. 2. Schematic figure of the non-parametric temporal pose estimation framework. Given an image observation $\mathbf{y}_t$, a set of pose hypotheses $\mathbf{X}_t$ are drawn from the model. Each hypothesis is given a temporal likelihood based on consistency with the hypothesis in the previous frame. The final estimate is the pose associated with the largest probability.

be extracted from the image and is capable of generalizing over different objects. As we will show in the experimental section, this sufficiently models the correlation between hand and object allowing estimation in scenarios with severe occlusion.

Having acquired a low-dimensional efficient representation $\mathbf{y}$ of the image as described above, the database is completed by associating each image $\mathbf{y}_i$ with its corresponding pose parameters $\mathbf{x}_i$. The pose vector $\mathbf{x}$ is composed of the rotation matrix of the wrist w.r.t. the camera and the sines of the joint angles of the hand.

5. Inference

As shown in Eq. (2), the conditional probability density over hand pose $\mathbf{x}_t$ is factorized into two different terms, an observation likelihood $p(\mathbf{x}_t|\mathbf{y}_t)$ and a temporal consistency model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$. Below we discuss these two models in more detail, and show how the pose $\mathbf{x}_t$ is estimated from the observation $\mathbf{y}_t$ using the implicit database model.

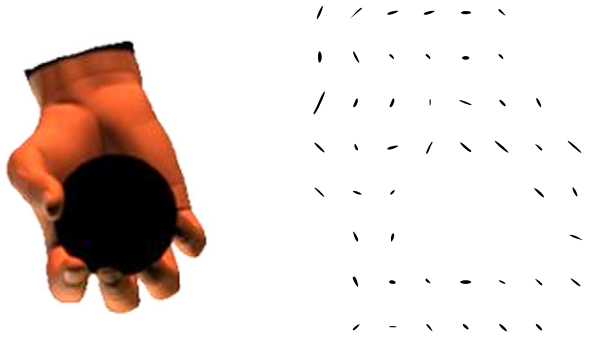


Fig. 3. The left image shows an example from the database. The right image shows the associated image feature descriptor $\mathbf{y}$. Prior to extracting the feature descriptor the object is segmented from the image, resulting in a "hole" at the corresponding position in the descriptor. This encodes the correlation between pose and object in a more robust manner compared to if the internal edges of the object would also contribute to the descriptor.

5.1. Observation

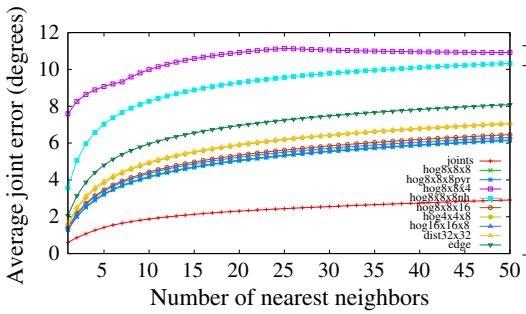
The pdf $p(\mathbf{x}_t|\mathbf{y}_t)$ is approximated by indexing into the database of hand poses using the image representation $\mathbf{y}_t$, and retrieving the nearest neighbors in the space spanned by the set of database features $\mathbf{Y}$. Due to the size of the database, an exact NN approach would be too computationally intensive. We therefore consider approximate methods. We compare Locality Sensitive Hashing (LSH) [31] and Fast Library for Approximate Nearest Neighbors (FLANN) [32], see Fig. 5, and decide to use LSH in our experiments as it shows an attractive trade-off between computational complexity and prediction accuracy.

LSH projects the feature space into multiple hash tables. The hash tables are designed so that if two feature vectors are close in feature space, their correspondent hashes are the same (or at least similar in Multi-probe LSH [31]). The parameters required by this algorithm are the number of hash tables to build L and the number of nearby hashes to probe T. The rest of the parameters are optimized offline for a required percentage of true K-nearest neighbors R. We set those values to $L = 10$, $T = 50$ and $R = 95\%$ empirically. Each LSH query $\mathbf{y}_t$ returns an approximation to the K nearest neighbors (in our case $K = 500$). Each retrieved KNN $\mathbf{y}^i_t$ is associated with a weight w^i_t from a spherical Gaussian density,

$$w^i_t = \mathcal{N}\left(\mathbf{y}^i_t \middle| \mathbf{y}_t, \sigma_y \mathbf{I}\right), \tag{3}$$

with standard deviation σ_y is set by experimental evaluation. This encodes our belief that the image feature representation is locally smooth and reduces the effect of erroneous neighbors from the LSH algorithm.

Each image feature in the database, $\mathbf{y}^j$ is associated with a pose $\mathbf{x}^j$. The poses $\{\mathbf{x}^i_t\}$ corresponding to the NN $\{\mathbf{y}^i_t\}$ can thus be retrieved. Together with the weights, they form the set $\{\mathbf{x}^i_t, w^i_t\}$ which is a sampled non-parametric approximation of $p(\mathbf{x}_t|\mathbf{y}_t)$.



NN	joints	8|8|8	pyr	8|8|4	hognh	8|8|16	4|4|8	16|16|8	dist	edge
1	0.61	1.32	1.29	7.59	3.57	1.42	1.57	1.39	1.64	2.07
2	0.87	2.04	1.99	8.27	5.05	2.19	2.44	2.15	2.51	3.14
3	1.08	2.55	2.51	8.64	5.97	2.73	3.05	2.68	3.13	3.88
4	1.26	2.94	2.89	8.89	6.58	3.14	3.50	3.08	3.58	4.41
5	1.42	3.25	3.20	9.07	7.03	3.46	3.85	3.39	3.93	4.81
6	1.55	3.50	3.45	9.21	7.38	3.72	4.14	3.65	4.21	5.12
7	1.65	3.71	3.66	9.32	7.66	3.94	4.37	3.86	4.45	5.38
8	1.74	3.89	3.84	9.59	7.89	4.12	4.58	4.05	4.65	5.60
9	1.81	4.05	4.00	9.81	8.10	4.29	4.76	4.21	4.82	5.79
10	1.87	4.19	4.14	9.99	8.27	4.43	4.91	4.35	4.98	5.96

Fig. 4. Average (across nearest neighbors) of the mean pose error of non-parametric pose estimation using different image features. The curves show the cumulative Euclidean distance between the query pose and its nearest neighbor number 1–50 in the database. Joints is the ground truth error in pose space, acquired by taking the nearest neighbors in the pose space directly. This is a lower bound on the error and shows the density of our database. The curves hogA × A × B show the error when using HOGs with $A \times A$ non-overlapping cells and a histogram of B bins (see Fig. 3 for an example of an $8 \times 8 \times 8$ HOG). The suffix pyr indicates that the HOG feature includes lower resolution cells (1×1, 2×2, ...,$A \times A$). The suffix nh means normalized holes: the histogram is normalized to sum to one (i.e., removing information on how large part of the cell is covered by skin colored areas). The curve dist32×32 shows the error when images are represented by their distance transform subsampled to 32×32 pixels. The edge curve shows the error when using the chamfer distance between edge maps extracted from the images. The result indicates that an $8 \times 8 \times 8$ HOG with pyramidal resolution HOG gives the lowest error, but $8 \times 8 \times 8$ HOG provides very similar performance with lower dimensionality. It is also interesting to note the importance of a sufficient number of bins, as it shows the bad results obtained by $8 \times 8 \times 4$ HOG.

5.2. Temporal consistency

As described in Section 3, the temporal consistency constraint $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is modeled as a parametric function g. It is used as a conditional prior to reweight the sampled distribution $\{\mathbf{x}_t^i, w_t^i\}$ approximating $p(\mathbf{x}_t|\mathbf{y}_t)$.

We assume that our model is getting observations densely enough in time such that the trajectory with respect to both the pose and view spaces varies smoothly. The naïve modeling approach would thus be to penalize estimates by their deviation in pose space to the previous estimate $\hat{\mathbf{x}}_{t-1}$. This model implicitly assumes that the temporal likelihood distribution $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is uni-modal. The uni-modality assumption can introduce unnecessary errors in the prediction since $\hat{\mathbf{x}}_{t-1}$ might not be the best candidate due to ambiguities (several poses can share a similar appearance) or estimation errors. A more sensible approach is to make use of all the hypotheses $\mathbf{X}_{t-1} = \{\mathbf{x}_{t-1}^i, \mathbf{w}_{t-1}^{*i}\}$ in the previous time instance and propagate them through time. We can do so by modeling the conditional distribution $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ using a kernel density estimation (KDE) approach [33], where the density is modeled as a mixture of Gaussian kernels centered in $\mathbf{x}_{t-1}^i$ and weighted by $\mathbf{w}_{t-1}^{*i}$. This enables the propagation of a potentially multi-modal distribution in time, making the temporal model significantly more flexible and expressive, allowing us to represent temporary ambiguities, resolving them further ahead in time.

As we will show in Section 6, having a strong temporal model allows us to perform prediction in noisy scenarios where the image observations are uncertain.

6. Experiments

We perform three sets of experiments using the proposed method. First we compare our non-parametric approach to a baseline of other state-of-the-art regression algorithms. In order to make an evaluation in terms of a quantitative error this experiment is performed using synthetic data where the joint configuration is known. Synthetic data also allows us to control the amount of noise in the images. Both our method and the baseline methods are evaluated in terms of robustness towards noise in the image observations. In the second set of experiments we evaluate our method in a qualitative manner on synthetic sequences with added image noise. The third set of experiments is performed on challenging real-world sequences.

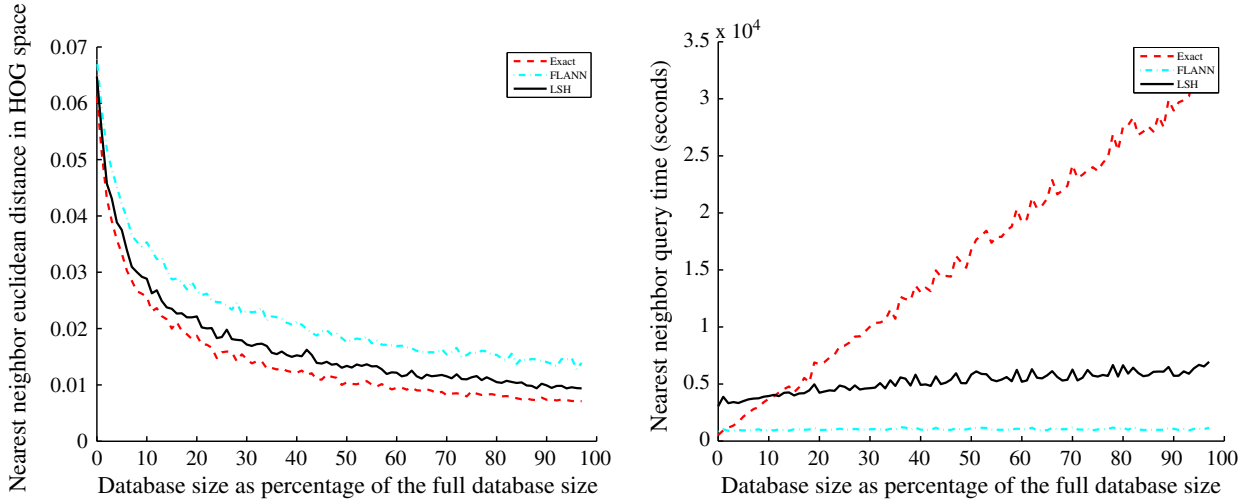


Fig. 5. The plot shows the prediction error (left) and average query time (right) as a function of database size (as percentage of the full database size) for finding the nearest neighbor in the database. 10% of the original database is set aside for testing, resulting in a full database of around 90,000 instances. Two approximate methods, LSH and FLANN, are compared with an exhaustive search as baseline. The left plot shows that LSH performs slightly better than FLANN in terms of accuracy. The right plot shows the query time increasing linearly for the exhaustive search while the approximate methods being sublinear, and FLANN being faster than LSH in absolute terms.

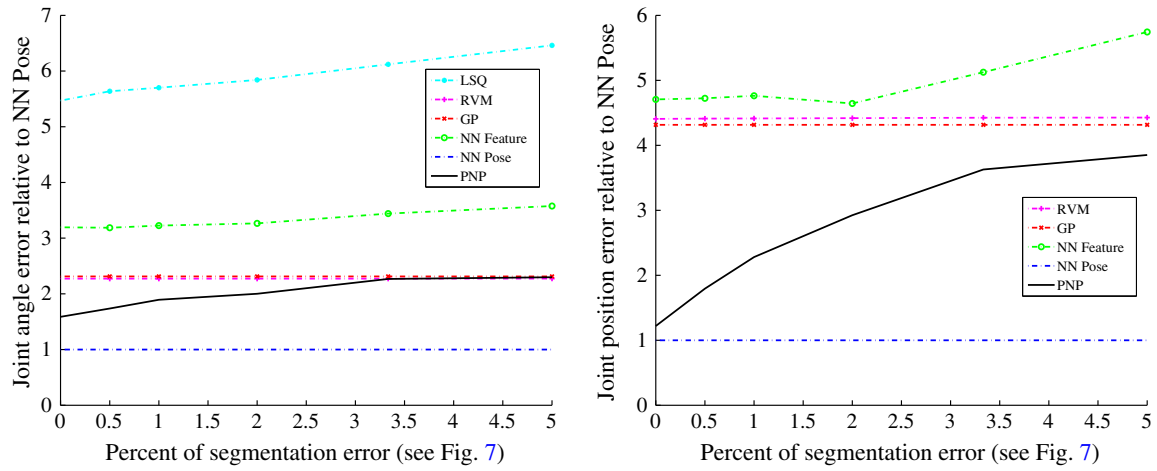


Fig. 6. Pose estimation using the non-parametric method (PNP) in comparison to three different regression techniques (LSQ, RVM, GP). As a baseline, the true nearest neighbor pose error (NN Pose) is shown, as well as the pose error of the nearest neighbor in feature space, not taking temporal information into regard (NN Feature). The plots show the average error with increasing segmentation noise, normalized with respect to the true nearest neighbor pose error. The error measure in the left plot is the Euclidean distance in the pose space spanned by **x**. The error measure in the right plot is proportional to the Euclidean distance in the space spanned by the 3D positions of all finger joints.

Videos of the real experiments can be seen at http://www.youtube.com/watch?v=RzenV-ma8Io.

6.1. Baseline

We compare our method to a set of regression models. In specific, we use Least Square Linear Regression (LSQ), the Relevance Vector Machine (RVM) [34] and Gaussian Process regression (GP) [35] to model the mapping from input features **y** to pose **x**, approximating the likelihood $p(\mathbf{x}|\mathbf{y})$ (no temporal information is included here). Each of these models has previously, with significant success, been applied to pose estimation [22,9,36] for both hands and full body pose.

All above models are based on a fundamental assumption that the mapping f^{-1} from image to pose takes functional form; LSQ assumes a linear form, while RVM and GP can model more flexible mappings. We compare these three methods to the suggested approach on four different synthetic sequences with varying degrees of added image noise, see Fig. 7. Neither the poses nor the objects in the test sequences are present in the database.

As can be seen in Fig. 6, left, the linear LSQ regression results in a very large error indicating that the relationship between feature and pose is inherently non-linear. The RVM and the GP are unable to model the mapping and do in fact always predict the same pose: the mean pose in the training data, irrespectable of image observation. In other words, this means that the appearance-to-pose mapping f^{-1} is under-constrained and does not take functional form. However, the non-parametric approaches are capable to model in such scenarios. From the results we can see that an exact nearest neighbor estimate in the feature space (without temporal information) results in a worse result compared to the mean pose distance in the data set, while our approach performs significantly better — also indicating that the mapping is non-unique. The dashed red line shows the results of an exact nearest neighbor in the pose space and is therefore a lower bound on the error of our method as it shows the resolution of the database.

The norm in joint space is not easily interpretable in terms of quality of the prediction as it does not respect the hierarchical structure of the hand, see Fig. 8. Therefore, the right plot of Fig. 6 shows the same mapping results, but with an error norm in terms of finger joint 3D positions. This shows even clearer how well our suggested method performs. With very little noise we are close to the exact NN lower bound, with increasing segmentation error asymptotically moving towards the mean.

Note that 5% error corresponds to a very weak segmentation, see Fig. 7. Further, our approach significantly outperforms the exact nearest neighbor in feature space (without temporal information). This clearly indicates how important temporal information is in order to disambiguate the pose.

To summarize, the results clearly show that the mapping from image features to pose is both highly non-linear and non-unique (multi-modal). This implies that it cannot be modeled using a functional approach.

6.2. Synthetic

In order to evaluate the qualitative performance of our method in a controlled scenario, we applied the model to image sequences with a controlled noise level. The results are visualized in Fig. 9.

The estimated pose over the two sequences is accurate while the associated object varies. This validates our assumption that objects generalize over pose and provide important contextual information.

6.3. Real sequences

In order to show the performance of our method in a real world manipulation scenario, we let three different subjects, two men and

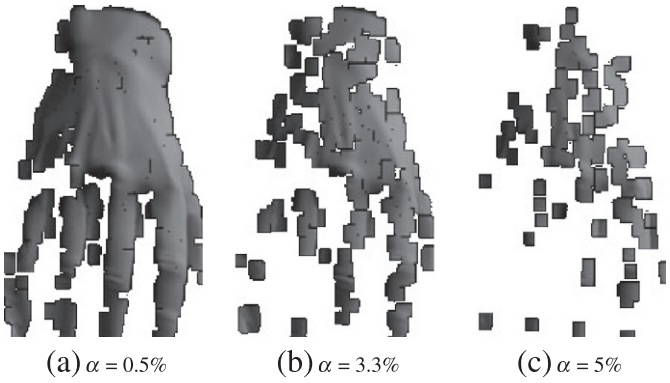


Fig. 7. Artificial corruption of the segmentation of the synthetic test data. The corruption is performed as follows: A partial segmentation is created by randomly removing α percentage of the pixels from the segmentation. The morphological operators of erosion and dilation then applied this partial segmentation in order to propagate the noise over the image. Examples of increasing segmentation noise are shown.

Fig. 8. Four different hand-poses are shown. The right-most image corresponds to the ground truth pose and the remaining images are estimates of the ground-truth. The estimates are ordered according to decreasing joint angle error. This clearly exemplifies how badly joint angle error corresponds to the quality of the estimate. This is because the norm in joint space assumes each dimension to contribute equally to the quality of the prediction. Therefore it does not reflect the hierarchical structure of the hand where error higher up in the chain (such as in the last two examples) effects the position of every joint further down the chain compared to the first prediction where the errors are concentrated closer to the finger tips.

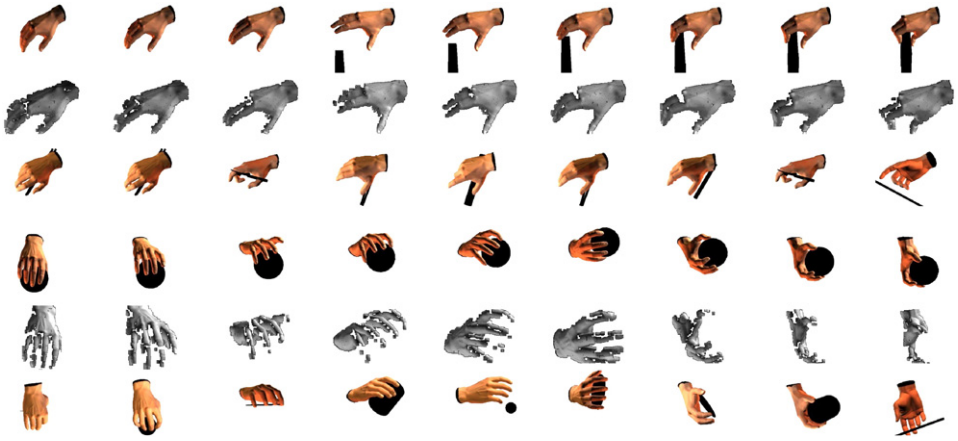


Fig. 9. Qualitative results of our approach applied to synthetic data. The top and the fourth row shows the ground truth pose, the second and the fifth rows show the segmentation from which the image features are computed. The segmentation has been corrupted by artificial noise with $\alpha = 0.5\%$ as explained in Fig. 7. The third and last rows show the corresponding predictions from our system. The two grasping sequences are applied to two different objects, in the first sequence a book and in the second a ball. We show the predicted hand-pose but also the object that is associated with the specific pose in the database.

one woman, manipulate three different objects. The objects are not contained within the model. The results are shown in Fig. 10.

As can be seen from the results, our model is capable of accurately predicting the pose of the hand. In each of the sequences the test hand shape and appearance is different from the database hand model, while there is no observable degradation in performance, showing that our model is robust to different hands. Further, as neither of the manipulated objects are represented in the model this further supports

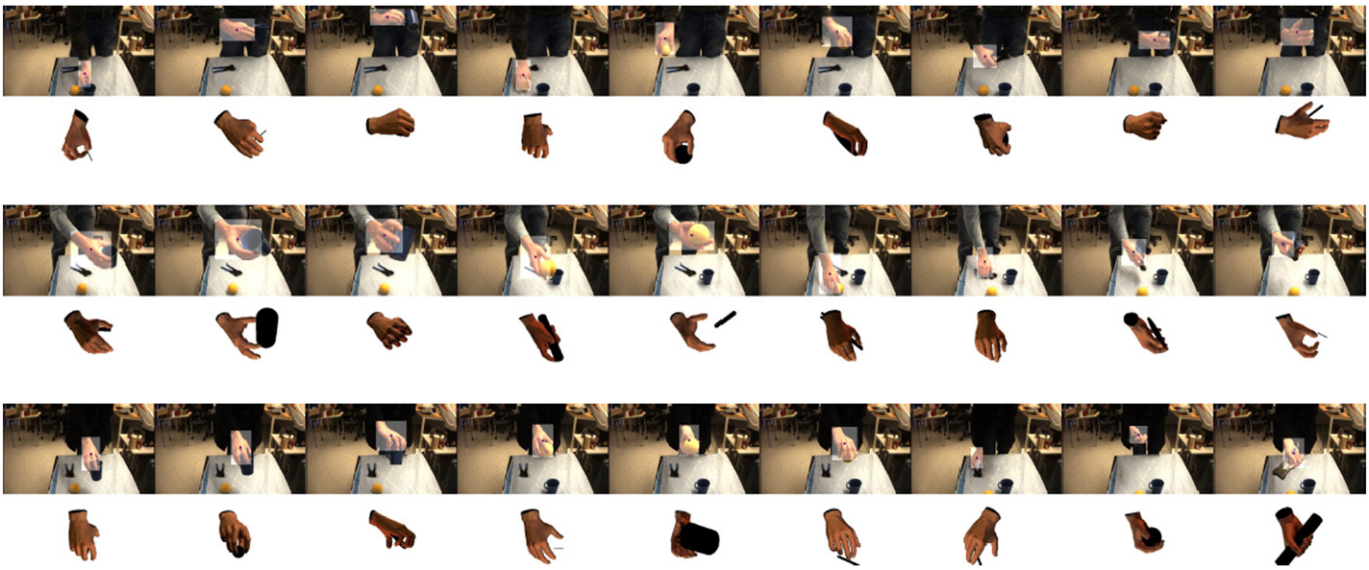


Fig. 10. Predictions of real world sequences. The three rows show three different sequences where different objects are manipulated by different humans. In the first and second sequences the subject is male while in the last one female. None of the objects exist in the database. The first, third and fifth rows show the input images with the skin detection window highlighted. The remaining rows show the associated predictions. As can be seen, the model correctly predicts the hand pose in each of the three different sequences.

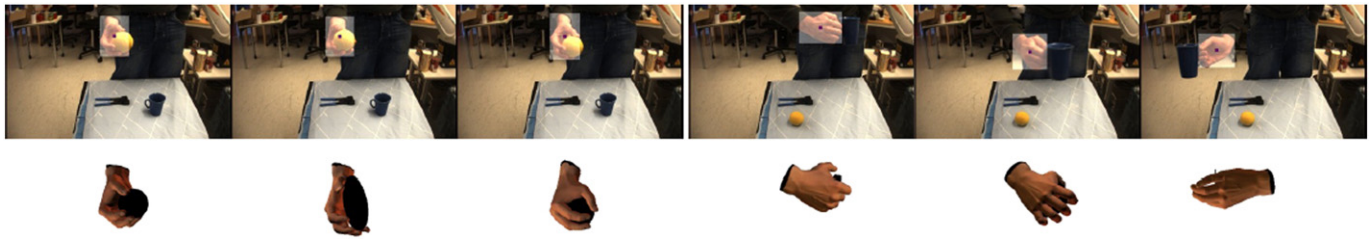


Fig. 11. The above sequences shows two challenging examples. In the left sequence a significant portion of the hand is occluded by the object. However, our proposed method still manages to correctly estimate the pose of the hand. This clearly shows the strength of jointly estimating the object and the pose rather than seeing them as independent. The right sequence is an example where the subject manipulates the objects in a rapid fashion in a highly non-linear manner. In such scenarios most dynamical models commonly applied in pose estimation will over smooth the solution or be unable to predict at all due to being fundamentally auto-regressive approaches. Our model correctly predicts the pose in the two first frames while the last estimate is erroneous. This error is an implication of the Markov one assumption in our temporal model which thereby is not capable of modeling inertia and therefore is unable to resolve the ambiguity in the image sequence.

the notion that grasps generalize over objects and that the objects' influence on the grasp provide important cues. This clearly shows that our system is capable of exploiting such information.

A large portion of the dynamical models that have been proposed to the problem of pose estimation are based on auto-regressive models [37], which assume that the trajectory in time takes functional form. Even though our dynamical model is parametric, it is based on the hypotheses from the non-parametric NN model. This means that it is considerably more flexible and can recover from bad estimates in situations where an auto-regressive model will fail. To highlight this strength we tested our model to a set of highly challenging sequences with fast non-linear motion and significant occlusion. This results in significant errors in the visual features. In Fig. 11 the results clearly show the strength of our approach, as it is able to track in such scenarios, and recover from errors which are difficult to avoid.

Further, we would like to highlight the efficiency of our algorithm. The method was implemented in C++ and runs at 10 frames/s on one of the cores of a four core 2.66 GHz Intel processor. Its speed makes it applicable in many different scenarios where pose estimation is an important source of information, and integratable with other computing-intensive algorithms.

7. Conclusions

We present an efficient non-parametric framework for full 3D hand pose estimation. We show through extensive experimentation that the proposed model is capable of predicting the pose in highly challenging scenarios corrupted by significant noise or with rapid motions. Further, our model is efficient and runs in real-time on standard hardware.

The fundamental contribution is a system capable of exploiting contextual information in the scene from the interaction between the hand and a potential object. We show how this information can be exploited in a robust manner, making our system capable of generalizing the pose over different objects. This enables the usage of a fast discriminative method to scenarios where only expensive generative methods previously would have been applicable. We employ a multi-modal temporal model, allowing us to resolve ambiguities through temporal consistency. Our model could easily be extended to simultaneously estimate both the hand pose and the object shape by appending the inference scheme with a smoothness term with respect to object.

In future work we would like to evaluate the possibility of exploiting a better pose representation. This would make it possible to even further strengthen the temporal model. In this paper we also assume that the observation model can be modeled using a spherical Gaussian; this encodes an assumption of equal importance of the joint angles. This is unlikely to be true why we would like to explore a likelihood model that better respects the correlation between quality of estimate in joint space. This could potentially allow us to use additional hypotheses for each estimate.

Another avenue of future work to investigate is the exploitation of RGB-D data, which would improve both the hand-background segmentation (currently based on skin color) and the feature representation of hand shape (currently HOG).

Finally, as noted in Section 2, generative and discriminative approaches have different merits. For applications requiring high accuracy, we therefore plan to run our discriminative hand pose estimator in parallel with a more accurate but less robust generative tracking method, using the discriminative estimates to (re)initialize the generative process.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.imavis.2013.04.002.

References

[1] J. Romero, H. Kjellström, D. Kragic, Hands in action: real-time 3D reconstruction of hands in interaction with objects, IEEE International Conference on Robotics and Automation, 2010.
[2] T.B. Moeslund, A. Hilton, V. Krüger, A survey of advances in computer vision-based human motion capture and analysis, Comp. Vision Image Underst. 104 (2–3) (2006) 90–126.
[3] A. Gupta, A. Kembhavi, L.S. Davis, Observing human-object interactions: using spatial and functional compatibility for recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (10) (2009) 1775–1789.
[4] G. Shakhnarovich, P. Viola, T. Darrell, Fast pose estimation with parameter-sensitive hashing, IEEE International Conference on Computer Vision, 2003.
[5] J. Shotton, A.W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, IEEE Conference on Computer Vision and Pattern Recognition, 2011.
[6] B. Yao, L. Fei-Fei, Modeling mutual context of object and human pose in human–object interaction activities, IEEE Conference on Computer Vision and Pattern Recognition, 2010.
[7] V. Athitsos, S. Sclaroff, Estimating 3D hand pose from a cluttered image, IEEE Conference on Computer Vision and Pattern Recognition, 2003.
[8] B.D.R. Stenger, A. Thayananthan, P.H.S. Torr, R. Cipolla, Model-based hand tracking using a hierarchical Bayesian filter, IEEE Trans. Pattern Anal. Mach. Intell. 28 (9) (2006) 1372–1384.
[9] T.E. de Campos, D.W. Murray, Regression-based hand pose estimation from multiple cameras, IEEE Conference on Computer Vision and Pattern Recognition, 2006.
[10] A. Thayananthan, R. Navaratnam, B. Stenger, P.H.S. Torr, R. Cipolla, Pose estimation and tracking using multivariate regression, Pattern Recognit. Lett. 29 (9) (2008) 1302–1310.
[11] R.Y. Wang, J. Popovic, Real-time hand-tracking with a color glove, ACM Trans. Graph 28 (3) (2009).
[12] H. Hamer, K. Schindler, E. Koller-Meier, L. Van Gool, Tracking a hand manipulating an object, IEEE International Conference on Computer Vision, 2009.
[13] I. Oikonomidis, N. Kyriazis, A.A. Argyros, Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints, IEEE International Conference on Computer Vision, 2011.
[14] H. Kjellström, J. Romero, D. Kragic, Visual object–action recognition: inferring object affordances from human demonstration, Comp. Vision Image Underst. 115 (1) (2011) 81–90.
[15] T. Feix, R. Pawlik, H. Schmiedmayer, J. Romero, D. Kragic, A comprehensive grasp taxonomy, RSS Workshop on Understanding the Human Hand for Advancing Robotic Manipulation, 2009.

[16] A. Erol, G.N. Bebis, M. Nicolescu, R.D. Boyle, X. Twombly, Vision-based hand pose estimation: a review, Comp. Vision Image Underst. 108 (2007) 52–73.
[17] R. Wang, S. Paris, J. Popovic, 6D hands: markerless hand-tracking for computer aided design, ACM Symposium on User Interface Software and Technology, 2011.
[18] A. Agarwal, B. Triggs, 3D human pose from silhouettes by relevance vector regression, IEEE Conference on Computer Vision and, Pattern Recognition, 2004, pp. 882–888.
[19] J. Romero, H. Kjellström, D. Kragic, Monocular real-time 3D articulated hand pose estimation, IEEE-RAS International Conference on Humanoid Robots, 2009.
[20] C.H. Ek, P.H.S. Torr, N.D. Lawrence, Gaussian process latent variable models for human pose estimation, Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, 2007.
[21] R. Urtasun, D.J. Fleet, P. Fua, 3D people tracking with Gaussian process dynamical models, IEEE Conference on Computer Vision and Pattern Recognition, 2006.
[22] A. Agarwal, B. Triggs, Recovering 3D human pose from monocular images, IEEE Trans. Pattern Anal. Mach. Intell. 28 (1) (2006) 44–58.
[23] S. Ekvall, D. Kragic, Interactive grasp learning based on human demonstration, IEEE International Conference on Robotics and Automation, 2004.
[24] S. Ekvall, D. Kragic, Grasp recognition for programming by demonstration tasks, IEEE International Conference on Robotics and Automation, 2005.
[25] A.A. Argyros, M.I.A. Lourakis, Real time tracking of multiple skin-colored objects with a possibly moving camera, European Conference on Computer Vision, 2004.
[26] A. Kanaujia, C. Sminchisescu, D.N. Metaxas, Semi-supervised hierarchical models for 3D human pose reconstruction, IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[27] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[28] G. Mori, S. Belongie, J. Malik, Efficient shape matching using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 27 (11) (2005) 1832–1837.
[29] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, IEEE Conference on Computer Vision and Pattern Recognition, 2005.
[30] G. Borgefors, Distance transformations in digital images, Comput. Vis. Graph. Image Process. 34 (3) (1986) 344–371.
[31] W. Dong, Z. Wang, M. Charikar, K. Li, Efficiently matching sets of features with random histograms, ACM Multimedia, 2008.
[32] M. Muja, FLANN, fast library for approximate nearest neighbors, http://mloss.org/software/view/143/. 2009.
[33] V. Morariu, B. Srinivasan, V. Raykar, R. Duraiswami, L. Davis, Automatic online tuning for fast Gaussian summation, Neural Information Processing Systems, 2008.
[34] M.E. Tipping, Sparse Bayesian learning and the relevance vector machine, J. Mach. Learn. Res. 1 (2001) 211–244.
[35] C.E. Rasmussen, Gaussian processes in machine learning, Advanced Lectures on Machine Learning: ML Summer Schools, 2003. (Canberra, Australia, Tübingen, Germany).
[36] X. Zhao, H. Ning, Y. Liu, T. Huang, Discriminative estimation of 3D human pose using Gaussian processes, IAPR International Conference on Image Processing, 2008.
[37] J. Wang, D. Fleet, A. Hertzmann, Gaussian process dynamical models for human motion, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2) (2008) 283–298.