Hybrid One-Shot 3D Hand Pose Estimation by Exploiting Uncertainties

Georg Poier[1]
poier@icg.tugraz.at
Konstantinos Roditakis[2,3]
croditak@ics.forth.gr
Samuel Schulter[1]
schulter@icg.tugraz.at
Damien Michel[2]
michel@ics.forth.gr
Horst Bischof[1]
bischof@icg.tugraz.at
Antonis A. Argyros[2,3]
argyros@ics.forth.gr

[1] Institute for Computer Graphics and Vision
Graz University of Technology
Graz, Austria

[2] Institute of Computer Science
FORTH
Heraklion, Greece

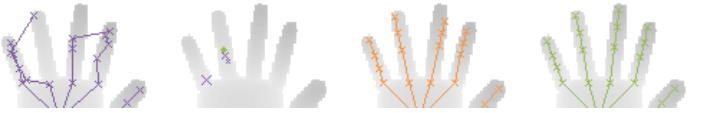
[3] Computer Science Department
University of Crete
Heraklion, Greece

Abstract

Model-based approaches to 3D hand tracking have been shown to perform well in a wide range of scenarios. However, they require initialisation and cannot recover easily from tracking failures that occur due to fast hand motions. Data-driven approaches, on the other hand, can quickly deliver a solution, but the results often suffer from lower accuracy or missing anatomical validity compared to those obtained from model-based approaches. In this work we propose a hybrid approach for hand pose estimation from a single depth image. First, a learned regressor is employed to deliver multiple initial hypotheses for the 3D position of each hand joint. Subsequently, the kinematic parameters of a 3D hand model are found by deliberately exploiting the inherent uncertainty of the inferred joint proposals. This way, the method provides anatomically valid and accurate solutions without requiring manual initialisation or suffering from track losses. Quantitative results on several standard datasets demonstrate that the proposed method outperforms state-of-the-art representatives of the model-based, data-driven and hybrid paradigms.

1 Introduction

The accurate estimation and tracking of hand articulations provides the basis for many applications like human computer interaction, activity analysis, and sign language recognition. Hand pose estimation, despite its long history [11, 32, 35, 46], has attracted more and more interest from the computer vision community in recent years [27, 33, 37, 48]. This is partly driven by the fact that low-cost sensors can now provide reliable depth information.



(a) (b) (c) (d)

Figure 1: (a) A learned joint regressor might fail to recover the pose of a hand due to ambiguities or lack of training data. (b) We make use of the inherent uncertainty of a regressor by enforcing it to generate multiple proposals. The crosses show the top three proposals for the proximal interphalangeal joint of the ring finger for which the corresponding ground truth position is drawn in green. The marker size of the proposals corresponds to degree of confidence. (c) A subsequent model-based optimisation procedure exploits these proposals to estimate the true pose. (d) The ground truth for this particular example. The same colours are used for the corresponding results throughout the work.

Most current approaches to tracking of hand articulations can be roughly categorized into model-based and data-driven schemes. In model-based schemes [10, 29, 30, 46] an underlying 3D hand model is used to render pose hypotheses, which are subsequently compared to the observations retrieved from the sensor. Since it is infeasible to search the whole range of possible hand poses, these methods rely on an initialisation which already needs to be close to the true solution. Typically, the solution from the previous frame is used for initialisation, which leads to problems in the case of very fast hand movements or dropped frames. Hence, subsequent tracking failures are hard to recover from.

On the other hand, data-driven schemes learn the mapping from specific appearances to hand poses from training data [22, 40]. During testing, they usually infer joint locations independently from each other. In this way, the complex dependencies do not need to be modelled. However, the results are not constrained by hand anatomy or physics. Thus, the obtained pose estimates might be wrong or even impossible. Another issue for these approaches is that employing enough training data to densely cover the whole pose space is infeasible, because of the highly articulated nature of the human hand and the fact that the space of possible hand poses grows exponentially with the number of joints.

In this work we propose a hybrid method with both data-driven and model-based elements that inherits the advantages of both paradigms. Very recently, this direction has attracted considerable research (*e.g.* [2, 33, 39, 43]), which has been inspired in parts by several relevant and successful approaches on human pose estimation. In general, there are two types of hybrid approaches. Either these approaches try to find a solution using the model-based approach first, and only consider the data-driven approach for failure cases [16, 45], or they obtain (an) initial pose(s) based on the data-driven approach, and then validate and/or locally optimise the pose(s) using a model-based approach [3, 21, 49]. While the chosen optimisation scheme, failure detection and runtime performance are critical issues for the former approaches, the latter ones naturally avoid the computationally difficult problem of global optimisation and do not get stuck to local minima. However, several of the latter approaches either only validate the hypotheses generated by the data-driven approach using the model-based method [35, 47], or do not consider anatomic constraints. Thus, the resulting poses might be invalid [3, 35, 49]. Moreover, for the task of 3D hand pose estimation, inherent difficulties like the substantial similarities between individual fingers and the very fast movements or complex finger interactions cause ambiguities and uncertainties which are often disregarded by previous works.

1.1 Exploiting Uncertainties

For deriving more informed decisions under uncertainty, Graphical Models have been proven very effective [12, 14, 24]. Hence, they have been extensively used in computer vision literature [13, 17, 20]. Besides being applied in a dense manner [15, 20, 25], sparse graphical models have become increasingly popular for tasks like object detection or pose estimation [9, 13, 19, 46]. Despite their effectiveness, a naive implementation would lack efficiency due to the complex interactions which need to be modelled. To this end, approximations of the underlying distributions have enabled efficient inference [13, 25, 26].

To exploit the inherent uncertainties in the task of 3D hand pose estimation, we first need to capture them. For this, we adopt successful work on body pose estimation [18, 38] to train a regressor that is able to generate a distribution of location proposals for each joint. We input this distribution to a subsequent optimisation procedure.

Within the optimisation procedure we can then exploit the uncertainties, which are implicitly captured by the proposal distribution. To do this efficiently we employ an approximation of the full distribution upon which a graphical model operates. The optimisation procedure considers multiple entirely different solutions for the global pose configuration, capturing the uncertainty of preceding processing steps. In this way, the regressor does not need to be perfectly accurate on its own but should only deliver a set of likely joint positions, which are subsequently refined. This also attenuates the need for a complete training database densely covering the whole pose space. Additionally, the whole process operates in 3D and is thus able to infer correct joint locations even in the case of occlusions and missing depth information. Moreover, optimisation not only exploits the uncertainties, but also finds an anatomically valid hand pose, similar to inverse kinematics (see Fig. 1).

1.2 Related Work

In previous attempts to apply hybrid approaches specifically to hand pose estimation, the joint proposals provided by the data-driven approach are often refined by adding penalties to anatomically implausible joint locations [32, 39]. However, the obtained hand poses can still be invalid since the refinement performs only a selection of the most plausible joint proposals and/or refines only some of the provided proposals. Hence, the approaches fail if all proposals for a single joint are inaccurate (*e.g.*, due to occlusions), or if the proposals are uncertain for many of the joints.

In contrast, our method introduces new joint positions, which respect anatomic constraints and, simultaneously, best fit the joint positions proposed by the data-driven approach in a global manner. Moreover, the proposed method does not rely on finding proposals of high confidence, but incorporates the approximated proposal distributions for each joint to find the anatomically valid hand pose which explains them best.

Another strand of research combines salient point detection with model-based optimisation [4, 33, 44]. However, these approaches rely on detection of specific landmarks and will fail in situations where the landmarks (usually the finger tips or nails) are not clearly visible. In contrast, we do not rely on any landmark to be visible, but take a more holistic view considering the whole hand. Thus, our approach is robust to occlusion of specific landmarks.

Also noteworthy is an approach to 2D body pose estimation [8, 9]. This DPM based work focuses on improving the unaries provided by a Random Forest. In contrast to their work, we use hard constraints on the graphical model, which is enabled by defining it in 3D. However, their ideas for improving the unaries could potentially be applied to our task too.

Probably most closely related to our method is the approach from Tompson *et al.* [43]. This approach uses a deep convolutional network to infer the most likely positions of some predefined points on the hand which are then optimised by a model similar to ours. Despite employing depth information, regression is performed solely in 2D, disregarding occlusions or "holes" in the depth map. Moreover, in contrast to our approach, they rely on a single best location to fit the model, which ignores the uncertainty in the regression.

The basic building blocks of our method are similar to those of other hybrid approaches to hand pose estimation: a discriminative, data-driven method that generates likely joint positions (as in [32, 39]), and a generative, model-based optimisation method that refines the initial solution (as in [33, 43]). However, the way we combine these two components is shown to outperform competing approaches. Another key difference that makes our method distinct from any other method we are aware of is that the optimisation component has access to internal information of the data-driven component. It can thus make more informed decisions under the given uncertainty, which again yields significantly improved results.

2 Method

In this section we present the two building blocks which make up the proposed method. We use a discriminative regressor, which generates an approximation of the proposal distribution (Sec. 2.1). This distribution can be effectively transformed to anatomically valid pose hypotheses using the model-based optimisation procedure described in Sec. 2.2.

2.1 Joint Regression

For the generation of an approximated proposal distribution we build upon the prominent approach of Girshick *et al.* [18, 38]. This approach has been shown to work well for real world applications on body pose estimation [38], and has also been previously adapted for hand pose estimation [39]. The approach relies on Random Forests [2, 5, 7] to infer a 3D distribution of likely hand joint locations[1]. We briefly describe the training and testing procedures as applied in this work since its internal information is later exploited during optimisation (Sec. 2.2). For more details the interested reader is referred to the related work.

Training: For our task we follow a part based approach to learn a mapping $\mathcal{F}:\mathcal{X}\rightarrow\mathcal{Y}$. An input sample $\mathbf{x}\in\mathcal{X}=\mathbb{R}^D$ represents the local appearance of an image patch around a foreground pixel from which we want to infer the 3D locations of J joints, *i.e.*, $\mathcal{Y}=\mathbb{R}^{3\times J}$. A training sample is formed by associating the image patch with a part label and corresponding offset vectors from the patch centre location to each of the joint positions. Hence, the training set $\mathcal{L}=\{(\mathbf{x}_i,c_i,\mathbf{o}_i)\}\subseteq\mathcal{X}\times\mathcal{C}\times\mathcal{O}$, where $c_i\in\mathcal{C}=\{1,\ldots,J\}$ denotes the class label of the joint which is closest to the location of $\mathbf{x}_i$, and $\mathbf{o}_i\in\mathcal{O}=\mathbb{R}^{3\times J}$ denotes the set of 3D offset vectors. The training data is recursively split by each tree individually, until the maximum depth of a tree (23 in this work) is reached or less than a minimum number of samples (40) arrives at a node. For the experiments we fixed the number of trees to three.

Inspired by [36], we sub-sample the data arrived at a node for split selection. This not only speeds up the training process and enforces de-correlation between the trees, but also implicitly accounts for the different number of samples which are extracted per class by drawing a balanced sub-sample. The learned split functions are based upon the same simple

[1] By joint locations we refer to any annotated positions defining the pose of a hand (most often these are joints)

depth features used in [38]. Finally, to generate the prediction models at the leaves, mean-shift [6] is applied to the collected offset distributions for each joint [38].

Evaluation: During test time we start with an empty set of proposals $\mathcal{P}_j = \emptyset$ for each joint j. Image patches are sampled densely from the foreground region of the depth image and are passed down through each tree of the forest. Arriving at leaf l, the 3D centre position of a test sample $\mathbf{x}$ is offset by the stored offset vectors $\mathcal{O}_j^{(l(\mathbf{x}))}$ for each joint j to form sets of proposals $\{\mathcal{P}_j^{(l(\mathbf{x}))}\}$ which are added to the current sets: $\mathcal{P}_j \leftarrow \mathcal{P}_j \cup \mathcal{P}_j^{(l(\mathbf{x}))}$. Following [38] we keep only a reduced set of top confident proposals for each joint $\tilde{\mathcal{P}}_j \subseteq \mathcal{P}_j$ and, subsequently, perform mean-shift on those to extract the k top modes $\mathcal{P}_j^{(m)}$. We thus end up with at most $k \times J$ final proposals, each associated with a confidence score. We base the score on the number of initial proposals supporting the mode [38]. Importantly, this set of proposals and confidences approximates the proposal distribution for each joint.

2.2 Model-based Optimisation

Using the discriminative Random Forest based method described above, inference of the individual joint proposals is completely independent from the other joints. While, in this way, the complex dependencies do not need to be modelled, the resulting proposals are not necessarily compatible with anatomical constraints.

In order to obtain a valid pose we employ a predefined 3D model of a hand. We use a model with 26 degrees of freedom (DoFs). In this model the global pose of the hand, *i.e.*, position and orientation, has six DoFs, and each of the five fingers is specified by four more. These four parameters per finger encode angles where the base joint of each finger is assigned two DoFs and the two remaining hinge joints are each assigned one DoF. During optimisation these DoFs are constrained based on anatomical studies [1, 28] which avoids impossible configurations. Since a quaternion representation is used for the global orientation, the 26 DoFs are modelled by 27 parameters.

While the used hand model is, in principle, similar to what is used in related work on hand pose estimation and tracking (*e.g.*, [31, 33]), our model only specifies the joint positions instead of specifically designed geometric primitives [31, 33], or even a complete mesh [23, 37, 42]. As pointed out later in this section, this has very important implications on the computational performance of the optimisation process.

Having defined a hand model, the goal is to find the 27 model parameters which best describe the modes $\mathcal{P}^{(m)}$ of the joint proposals, obtained from the regression forest. To this end, the objective function $E(\mathcal{P}^{(m)}, \mathbf{h})$ judges the quality of any hypothesized parameter set $\mathbf{h} \in \mathbb{R}^{27}$. More specifically, given a function $\delta_j(\mathbf{h})$ which extracts the position of joint j from hypothesis $\mathbf{h}$, the objective is formulated as:

$$E(\mathcal{P}, \mathbf{h}) = \sum_{j=1}^{J} \max_r \left(w_{j_r} \left(1 - d_{j_r}^2 \right) \right), \tag{1}$$

where

$$d_{j_r} = \min \left(1, \frac{\left\| \mathbf{p}_{j_r} - \delta_j(\mathbf{h}) \right\|_2}{d_{max}} \right). \tag{2}$$

Here $\|.\|_2$ is the L^2 norm of the argument, d_{max} is the clamping distance, $\mathbf{p}_{j_r} \in \mathcal{P}$ denotes the r-th proposal for joint j, and w_{j_r} is the normalized confidence so that it exhibits the properties

of a probability. Intuitively, the objective enforces those proposals to be selected which – together – best form an anatomically valid pose. This selection is guided by the confidence of each proposal. Moreover, by considering all the top modes of the proposal distribution for a joint, we overcome the problem of outlier modes (*e.g.*, proposals for the wrong finger). The model will simply converge to joint positions close to those modes which best fit into the overall model. This is achieved by optimising the objective for the best parameter set $\mathbf{h}^\star$:

$$\mathbf{h}^\star \triangleq \underset{\mathbf{h}}{\operatorname{argmax}}\, E\left(\mathcal{P}^{(m)}, \mathbf{h}\right). \tag{3}$$

It is important to note that the definition of the objective function is based on a small number of 3D distances. As a result, no 3D hand model rendering is required to evaluate the objective function as, *e.g.* in [31, 37]. On the contrary, it can be computed very efficiently on conventional (*i.e.*, CPU) processors.

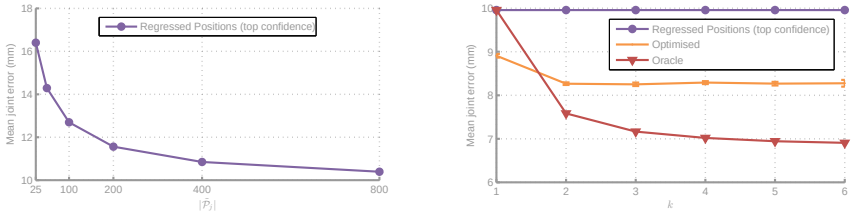
The objective function is optimised by Particle Swarm Optimisation (PSO) [21] as employed in many other works on 3D hand pose estimation [27, 31, 33, 37, 43]. PSO performs optimisation by evolving a number of particles (solutions) that evolve in parallel over a number of generations (iterations). If not stated otherwise, an overall number of 50 generations turned out to be sufficient for our experiments. The incorporated randomness, which is introduced during the initialisation of the particles as well as during their evolution, makes the method well suited for the employed non-convex, non-smooth objective function.

Stepwise optimisation: Any search space grows exponentially with the number of its dimensions. Hence, decomposing the space into non-overlapping sub-spaces offers the possibility of a very significant speed-up. We exploit this for hand pose estimation based on the observation that, given the global orientation of the hand, the fingers can move almost independently of each other. This independence together with the performed mapping of the proposals to specific joints allows us to split the optimisation problem into sub-problems of lower dimensionality. In line with this, we treat the problem of finding the best 27 parameters as six sub-problems, where we first optimise for the 7 parameters specifying the global pose of the palm, and subsequently individually optimise for the 4 parameters of each finger.

3 Experiments

We prove the applicability of the proposed method by means of several experiments on synthetic as well as real world data. A crucial issue for benchmark datasets is the availability of ground truth. However, accurate 3D annotations for real data are not easily obtained, especially for articulated self occluding objects like the human hand. To overcome this issue, we employ synthetic data in addition to real data. To generate a synthetic sequence which resembles natural movements, a hand is tracked with the method described in [31] using the publicly available implementation[2] and with a very high computational budget. We then render depth maps from the resulting poses and utilize the renderings as the test sequence. In this synthetic sequence the hand performs various finger articulations and typical motions like counting, pinching and grasping. While the performed movements are slow to ensure that the tracker does not get lost, we afterwards sampled every 5th frame from the sequence to simulate a more natural speed of movements. This sequence is referred to as *TrackSeq*.

[2] publicly available at http://cvrlcode.ics.forth.gr/handtracking/



(a) (b)

Figure 2: (a) Mean joint localization error on ICVL dataset as a function of the number of proposals, $\left|\tilde{\mathcal{P}}_j\right|$, which are input to mean-shift to extract the final joint proposals at test time. (b) Mean joint localization error on the *TrackSeq* sequence as a function of the number of top proposals, k. For *Optimised* the error bars show the standard deviation over multiple runs.

For producing training data, we first defined 4 different articulations per finger. All 1024 combinations of these articulations were used as an initial set of poses. These poses were then rendered under 7 different viewpoints to create the full train set[3] of 7168 poses.

For experiments with real data we employ the *ICVL Hand Posture Dataset*[4] and *NYU Hand Pose Dataset*[5], where we use the available training and test data as is. The ICVL dataset was acquired using the Intel Creative *Time-of-Flight* (ToF) camera and includes a training set with roughly 330K images and two test sequences with 702 and 894 images, respectively. The two test sequences show a hand facing towards the camera performing various finger articulations in very fast succession. The NYU dataset was acquired using the Kinect RGB-D camera and includes a training set with roughly 73K images and a test set capturing two actors and consisting of 8252 images (2440 and 5812, resp.). However, neither our method, nor that of Tompson *et al*. [43], who published the dataset, can yield any meaningful result for the second actor. This is because there is only a single actor included in the training set and the hand of the second actor in the test set differs significantly. Thus, we only compare on the sequence of the first actor.

3.1 Influence of Major Processing Steps

Size of proposal sets $\tilde{\mathcal{P}}_j$: An important parameter of our method is the number of proposals, $\left|\tilde{\mathcal{P}}_j\right|$, which are input to mean-shift at test time (see Sec. 2.1). This is especially interesting since mean-shift is responsible for summarizing the proposal distribution that we want to exploit during optimisation. As can be seen from Fig. 2(a), the gain in accuracy becomes smaller for a higher number of proposals. This enables us to find a good trade-off between speed and accuracy. In our case, we fix $\left|\tilde{\mathcal{P}}_j\right| = 200$ for all other experiments. Another interesting observation is that accuracy seems to level off much later than reported for body pose estimation in [38]. We hypothesize that this difference is due to the higher variation of hand poses compared to body poses within the respective datasets. In any case, it further advocates the specific consideration of the uncertainty as proposed in this work.

Number of final proposals per joint: The optimiser can efficiently exploit the inherent uncertainty of the regression process due to the approximation of the proposal distribution for each joint. Thus, we investigate the effect of the number of generated proposals k on

[3]Note, that rendering from a different viewpoint is equivalent to changing the orientation of the whole hand.
[4]publicly available at http://www.iis.ee.ic.ac.uk/~dtang/hand.html
[5]publicly available at http://cims.nyu.edu/~tompson/NYU_Hand_Pose_Dataset.htm

accuracy. Fig. 2(b) shows the error with respect to k. The *Oracle* always selects the proposal closest to the respective ground truth joint position. Obviously, the more proposals generated, the closer one of them will be to the ground truth. The error for the proposed method (*Optimised*) is higher because the solution has to respect the anatomical constraints of the hand. Interestingly, accuracy levels off for a small number of proposals (2-3) both for *Oracle* and for *Optimised*. Hence, the results show that utilizing a small number of proposals (together with confidences) instead of the full proposal distribution $\mathcal{P}^{(i)}$ is already very effective. For the other experiments we thus set $k = 3$. In fact this also limits the complexity of the objective function (Eq. (1)).

Furthermore, Fig. 2(b) indicates that the results of plain regression can already be improved by applying optimisation to the single top proposal for each joint. This improvement can be attributed to the anatomically valid solution induced by model-based optimisation. As also suggested by the results (for $k = 1$ and $k = 3$) in Fig. 3, a significant additional gain is achieved by providing the optimisation procedure with internal information about the uncertainty of the regressor, *i.e.*, multiple proposals per joint.

Stepwise optimisation: We investigate the effect of the stepwise optimisation procedure described in Sec. 2.2. For a meaningful comparison we fix the overall number of objective function evaluations (*i.e.*, the optimisation budget) for both approaches. We used 91 particles and generations when optimising all parameters together, whereas for the stagewise optimisation we assigned 64 particles and generations to the optimisation of position and orientation and 29 particles and generations to optimisation of each finger. The results are shown in Fig. 3(a) (orange curves). Despite the relatively high optimisation budget, the procedure does not converge when optimising all parameters together. That is, the stepwise optimisation procedure proves much more effective.

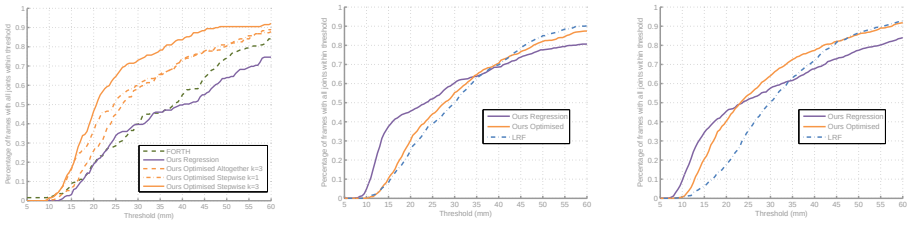
Runtime: Our current implementation of the Random Forest based regressor takes $\sim$33 msec on an Intel i7-4820K CPU to compute the joint proposals which are input to the optimiser. Optimisation itself obviously depends on the budget. For the current implementation 1000 objective function evaluations take $\sim$10 msec. If not stated otherwise, we fixed the number of objective evaluations to $\sim$3400 for all experiments for a good speed vs. accuracy trade-off. Note that all timings are given for our current prototype implementation, which is non-optimised and *single* threaded.

3.2 Comparison to the State of the Art

We compare our method to state of the art model-based (FORTH, [31]), data-driven (LRF, [40]) and hybrid (Tompson *et al.*, [43]) approaches. The LRF method has recently been shown to outperform other data-driven as well as model-based approaches [40]. Note that, all our results can be found on our website[6].

Proposed vs FORTH [31]: We evaluate our method on the *TrackSeq* test sequence. Fig. 3(a) compares the frame-based *success rate* over a number of distance thresholds. The frame-based success rate gives the ratio of frames in which all joints are estimated within a certain threshold to ground truth. The results show that our plain regressor performs similarly to [31] for the most interesting range of thresholds. However, after enforcing anatomic constraints by model-based optimisation, the proposed hybrid method is able to improve on these results by a large margin. This is achieved despite the fact that the sequence exhibits strong finger articulations, whereas the position and orientation do not change. Thus, the main reason for

[6] Results and other material can be found at http://lrs.icg.tugraz.at/research/hybridhape/



(a) (b) (c)

Figure 3: Success rate as a function of the distance threshold for (a) the *TrackSeq* test sequence, (b) the ICVL dataset *Sequence 1* and (c) the ICVL dataset *Sequence 2*. It shows the ratio of frames with *all* joints within a certain distance to the ground truth as a function of this distance. Results are compared to those from *FORTH* [31] and *LRF* [40].

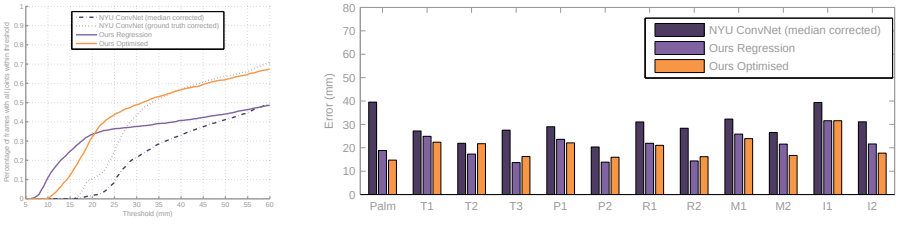
this gain is the improved estimation of finger articulations rather than the overall position and orientation estimation. This is also illustrated when solely considering the errors in finger tip localization, where the mean error for [31] is 19.5mm, while for our approach it is 11.8mm – an error reduction of about 40%.

Proposed vs LRF [40]: We compare to LRF [40] on the ICVL dataset. To ensure a fair comparison we use their results published online. As shown in Fig. 3(b) and 3(c) our regressor outperforms their results over most of the thresholds by a significant margin.

Unfortunately, we cannot fairly evaluate our model-based optimisation using the annotations provided with the ICVL dataset. This is because in the ground truth annotations, bone lengths[7] vary significantly between the frames of a single sequence; therefore they are not compatible with an anatomically valid 3D hand. However, the 26 DoFs hand model used in this work implicitly applies strict constraints on them. Fitting such a model will therefore always introduce additional errors when compared to the provided ground truth annotations. Nevertheless, our results appear more accurate or at least as accurate as those from LRF [40].

Proposed vs Tompson [43]: We compare to the hybrid approach of Tompson *et al.* [43] based on the NYU dataset. For this dataset the annotated positions do not actually correspond to joints, but to some specific locations on the hand. We used the suggested positions for evaluation [43] with the minor exception that we had to skip two of the three palm positions since the palm is only represented by a single position in our model. In addition, [43] provides only 2D locations, which are projected to 3D using the input depth. However, the input depth might be significantly distorted (*e.g.*, at holes). To correct inferred locations, which lie at positions with distorted depth, we augment them with the median depth of the inferred positions in the same frame. Fig. 4 shows the results and also compares to a variant where we correct the depth using the ground truth depth of the corresponding joints to show the theoretical upper bound for their approach. For our method the optimisation improves results particularly for larger distance thresholds since the optimisation mainly performs an error correction rather than improving inferred locations, which are already very close to ground truth. In addition, the difference between the annotation model and our model induces an error, especially for low thresholds. In spite of that, we observe that our method outperforms the approach of Tompson *et al.* [43] by a large margin – even if we correct their results by augmenting ground truth depth information.

[7] By bone lengths we refer to the distance between joint annotations connected by bones



(a) (b)

Figure 4: Results for the NYU dataset. (a) The ratio of frames where *all* inferred positions are within a certain distance from ground truth as a function of this distance. (b) The average error per joint. Results are compared to the hybrid approach of [43] (*NYU ConvNet*).

4 Conclusions

We proposed a hybrid approach for 3D hand pose estimation based on a single depth frame. A regression forest delivers several proposals for each hand joint position. Then, model-based optimisation is responsible for estimating the best fit of a 3D hand model to the joint proposals obtained through regression. Thus, optimisation exploits the inherent uncertainty of the data-driven regression. As a result, the proposed method delivers anatomically valid solutions (which most purely data-driven methods fail to provide) without unrecoverable track losses or a need for proper initialization (as happens with most purely model-based approaches). The proposed method has been shown to achieve state-of-the-art performance. This is proven by quantitative experiments on several datasets and in comparison with representative methods from all categories (model-based, data-driven, hybrid).

Acknowledgement We thank Axel Pinz for his efforts in reviewing the paper prior to submission, and gratefully acknowledge the support and inspiring discussions with Paschalis Panteleris, Nikolaos Kyriazis and Iason Oikonomidis from the CVRL lab at FORTH. This work has been supported by the European Union under the Seventh Framework Programme, projects 3D-PITOTI (ICT-2011-600545), FP7-IP-288533 Robohow and FP7-ICT-2011-9 WEARHAP, and by the Austrian Research Promotion Agency (FFG) under the FIT-IT Bridge programme, project TOFUSION (838513).

References

[1] Irene Albrecht, Jörg Haber, and Hans-Peter Seidel. Construction and animation of anatomically based human hand models. In *Proc. Eurographics Symposium on Computer Animation*, 2003.

[2] Yali Amit and Donald Geman. Randomized inquiries about shape; an application to handwritten digit recognition. Technical Report 401, Department of Statistics, University of Chicago, IL, 1994.

[3] Andreas Baak, Meinard Müller, Gaurav Bharaj, Hans-Peter Seidel, and Christian Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *Proc. IEEE Int'l Conf. on Computer Vision*, 2011.

[4] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *Proc. European Conf. on Computer Vision*, 2012.

[5] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[6] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[7] Antonio Criminisi, Jamie Shotton, and Ender Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7(2-3):81–227, 2012.

[8] Matthias Dantone, Jürgen Gall, Christian Leistner, and Luc Van Gool. Human pose estimation using body parts dependent joint regressors. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2013.

[9] Matthias Dantone, Juergen Gall, Christian Leistner, and Luc J. Van Gool. Body parts dependent joint regressors for human pose estimation in still images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36(11):2131–2143, 2014.

[10] Martin de La Gorce, David J. Fleet, and Nikos Paragios. Model-based 3d hand pose estimation from monocular video. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(9):1793–1805, 2011.

[11] Ali Erol, George Bebis, Mircea Nicolescu, Richard D. Boyle, and Xander Twombly. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding*, 108(1-2):52–73, 2007.

[12] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *Int'l Journal of Computer Vision*, 61(1):55–79, 2005.

[13] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[14] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. on Computers*, 22(1):67–92, 1973.

[15] Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto. Class segmentation and object localization with superpixel neighborhoods. In *Proc. IEEE Int'l Conf. on Computer Vision*, 2009.

[16] Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun. Real time motion capture using a single time-of-flight camera. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2010.

[17] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[18] Ross B. Girshick, Jamie Shotton, Pushmeet Kohli, Antonio Criminisi, and Andrew W. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *Proc. IEEE Int'l Conf. on Computer Vision*, 2011.

[19] Henning Hamer, Konrad Schindler, Esther Koller-Meier, and Luc Van Gool. Tracking a hand manipulating an object. In *Proc. IEEE Int'l Conf. on Computer Vision*, 2009.

[20] Xuming He, Richard S. Zemel, and Miguel Á. Carreira-Perpiñán. Multiscale conditional random fields for image labeling. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2004.

[21] James Kennedy and Russel C. Eberhart. Particle Swarm Optimization. In *Proc. IEEE Int'l Conf. on Neural Networks*, 1995.

[22] Cem Keskin, Furkan Kiraç, Yunus Emre Kara, and Lale Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In *Proc. European Conf. on Computer Vision*, 2012.

[23] Sameh Khamis, Jonathan Taylor, Jamie Shotton, Cem Keskin, Shahram Izadi, and Andrew Fitzgibbon. Learning an efficient model of hand shape variation from depth images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2015. (to be published).

[24] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models - Principles and Techniques*. MIT Press, 2009.

[25] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems*, 2011.

[26] Philipp Krähenbühl and Vladlen Koltun. Parameter learning and convergent inference for dense random fields. In *Proc. Int'l Conf. on Machine Learning*, 2013.

[27] Nikolaos Kyriazis and Antonis A. Argyros. Physically plausible 3d scene tracking: The single actor hypothesis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2013.

[28] John Y. Lin, Ying Wu, and Thomas S. Huang. Modeling the constraints of human hand motion. In *Proc. Workshop on Human Motion*, 2000.

[29] Stan Melax, Leonid Keselman, and Sterling Orsten. Dynamics based 3d skeletal hand tracking. In *Proc. Graphics Interface*, 2013.

[30] Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. Markerless and efficient 26-dof hand pose recovery. In *Proc. Asian Conf. on Computer Vision*, 2010.

[31] Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *Proc. British Machine Vision Conf.*, 2011.

[32] Rudra P. K. Poudel, Jose A. S. Fonseca, Jian J. Zhang, and Hammadi Nait-Charif. A unified framework for 3d hand tracking. In *Proc. ISVC*, 2013.

[33] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2014.

[34] James M Rehg and Takeo Kanade. Visual tracking of high dof articulated structures: an application to human hand tracking. In *Proc. European Conf. on Computer Vision*, 1994.

[35] Rómer Rosales and Stan Sclaroff. Combining generative and discriminative models in a framework for articulated pose estimation. *Int'l Journal of Computer Vision*, 67(3): 251–276, 2006.

[36] Samuel Schulter, Christian Leistner, Peter M. Roth, Luc Van Gool, and Horst Bischof. On-line Hough forests. In *Proc. British Machine Vision Conf.*, 2011.

[37] Toby Sharp, Cem Keskin, Duncan P. Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, Daniel Freedman, Pushmeet Kohli, Eyal Krupka, Andrew W. Fitzgibbon, and Shahram Izadi. Accurate, robust, and flexible real-time hand tracking. In *Proc. ACM Conf. on Human Factors in Computing Systems*, 2015.

[38] Jamie Shotton, Ross B. Girshick, Andrew W. Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, and Andrew Blake. Efficient human pose estimation from single depth images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(12):2821–2840, 2013.

[39] Danhang Tang, Tsz-Ho Yu, and Tae-Kyun Kim. Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *Proc. IEEE Int'l Conf. on Computer Vision*, 2013.

[40] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2014.

[41] Jonathan Taylor, Jamie Shotton, Toby Sharp, and Andrew W. Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.

[42] Jonathan Taylor, Richard V. Stebbing, Varun Ramakrishna, Cem Keskin, Jamie Shotton, Shahram Izadi, Aaron Hertzmann, and Andrew W. Fitzgibbon. User-specific hand modeling from monocular depth sequences. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2014.

[43] Jonathan Tompson, Murphy Stein, Yann LeCun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Trans. on Graphics*, 33(5):169:1–169:10, 2014.

[44] Dimitrios Tzionas, Abhilash Srikantha, Pablo Aponte, and Jürgen Gall. Capturing hand motion with an RGB-D sensor, fusing a generative model with salient points. In *Proc. German Conf. on Pattern Recognition*, 2014.

[45] Xiaolin K. Wei, Peizhao Zhang, and Jinxiang Chai. Accurate realtime full-body motion capture using a single depth camera. *ACM Trans. on Graphics*, 31(6):188:1–188:12, 2012.

[46] Ying Wu, John Y. Lin, and Thomas S. Huang. Capturing natural hand articulation. In *Proc. IEEE Int'l Conf. on Computer Vision*, 2001.

[47] Chi Xu and Li Cheng. Efficient hand pose estimation from a single depth image. In *Proc. IEEE Int'l Conf. on Computer Vision*, 2013.

[48] Chi Xu, Ashwin Nanjappa, Xiaowei Zhang, and Li Cheng. Estimate hand poses efficiently from single depth images. *Int'l Journal of Computer Vision*, 2015. (available under Open Access).

[49] Mao Ye, Xianwang Wang, Ruigang Yang, Liu Ren, and Marc Pollefeys. Accurate 3d pose estimation from a single depth image. In *Proc. IEEE Int'l Conf. on Computer Vision*, 2011.