Combining Discriminative and Model Based Approaches for Hand Pose Estimation

Philip Krejov, Andrew Gilbert and Richard Bowden
Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, United Kingdom

***Abstract*— In this paper we present an approach to hand pose estimation that combines both discriminative and model-based methods to overcome the limitations of each technique in isolation. A Randomised Decision Forests (RDF) is used to provide an initial estimate of the regions of the hand. This initial segmentation provides constraints to which a 3D model is fitted using Rigid Body Dynamics. Model fitting is guided using point to surface constraints which bind a kinematic model of the hand to the depth cloud using the segmentation of the discriminative approach. This combines the advantages of both techniques, reducing the training requirements for discriminative classification and simplifying the optimization process involved in model fitting by incorporating physical constraints from the segmentation. Our experiments on two challenging sequences show that this combined method outperforms the current state-of-the-art approach.**

I. Introduction

Estimation of hand pose has wide ranging applications covering areas such as gesture and sign language recognition, digital advertising, sterile computer use in operating theatres or home entertainment. The hands pose can be represented as a class of previously seen hand shapes or as the configuration of the hand, specified by joint positions and/or angles.

In the recent work of Tang [1] and Keskin [2], comparisons have been drawn between the challenge of hand pose and that of body pose estimation. This is because there are similar requirements for real time performance and accuracy in determining the configuration of an articulated object. However, due to limitations in the resolution and the small area of the finger tips, there are increased challenges in applying the methods of body pose estimation directly to hands.

Increased Variation. The large range of possible arm motion, results in a greater variety of global hand poses. The hand can be observed from a larger range of global rotations than bodies, which are typically limited to "feet on the floor" scenarios, as shown in figure 1. This additional Degree of Freedom (DoF) causes large variation in depth appearance, for similar joint configurations. Thus, extensive datasets (50,000+ images) are required to capture the huge variability of the hand.

Complex joint dependencies and range. The hand is comprised of a complex chain of kinematic relationships that can cause large scale occlusion and deformation, both of which create ambiguity when determining the hands pose. This range of flexibility also varies between users.

This work was supported by a EPSRC studentship and the EPSRC project, Learning to Recognise Dynamic Visual Content from Broadcast Footage (EP/I011811/1)

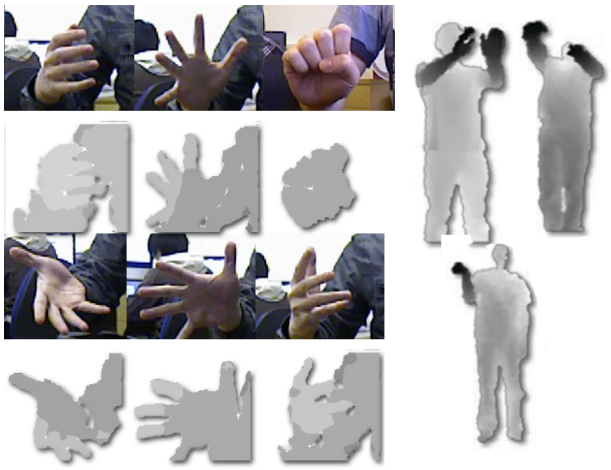


Fig. 1: Examples of typical hand shapes captured in depth showing noise and large amounts of global rotation. Body poses exhibit less global rotation and more depth detail.

Noisy hand data. Hands captured using depth can exhibit large amounts of contour noise and missing depth data. This noise is challenging to reproduce in synthetic data [3], and recover when using real data [4]. It is also challenging to obtain large quantities of accurate labelled ground truth data for machine learning approaches.

In solving hand pose estimation several criteria must be considered.

Preconditions to successful hand pose estimation

1) *Kinematic limitations* – manoeuvrability defined by the anatomical structure of the hand.
2) *Invalid self-intersection* – part of the hand entering the volume of another part constitutes an invalid pose.
3) *Temporal cohesion* – the understanding that in an image sequence the hand and its parts will transition from one location to another over the course of multiple frames.
4) *Depth observation* – Joints must reside within the observation of the hand unless part of the observation is missing.

This paper investigates a combined approach using both a global approximation and local optimisation, combining both discriminative and model based approaches. The global approximation provides a robust, coarse estimate, while the local optimisation refines a parametric model to fit the hands appearance. The global search is performed using a Randomised Decision Forest (RDF) trained using labelled depth

data. The local optimisation uses Rigid Body Dynamics and data driven constraints to efficiently model the hand. As the model is posed in a physics driven framework, tracking is handled implicitly by the simulation.

We discuss the benefit of using this combined optimisation method, and its application to hand pose estimation. Our evaluation shows that the use of a global estimator in combination with local refinement, improves on state of the art for estimation of hand joints in the dataset of Tang [1].

This paper is divided into the following sections. Section II discusses the field of hand pose estimation. Section III discusses our approach to solving hand pose. Evaluation is then performed against state of the art in Section III-B. Section V then provides our conclusions and discusses future work.

II. Related Work

In general, work to determine the pose of the hand can categorised as either model based, discriminative or shape analysis:

A. Model Inference

Model based methods attempt to determine the hand configuration using a generative model of the hand. This is commonly achieved with estimated pose parameters typically derived from the previous frames pose using an optimisation framework in an attempt to establish the model parameters through iterative refinement. Searching for a solution where the observed hand and rendered model converge in appearance provides a new estimate of the pose. Many approaches aim to reduce the computation, by reducing either the search space or parameter space. Using local encoding of the states of the finger joints, Sudderth [5] models kinematic and structural constraints, but do not account for loss of tracking. A probabilistic framework proposed by Stenger [6] aims to alleviate this by re-initialising after failure of tracking. A tree structure is used to carve the search space to avoid unlikely poses. Hamer [7] models the hand by parts which allows partial matching of the pose for heavily occluded hands grasping objects. To remove ambiguity due to edge or silhouette information, DeLaGorce [8] models the lighting and texture of the hand in a constrained fashion which is shown to improve accuracy. Furthermore, depth based approaches have been proposed as these are not effected by illumination changes. One such approach is that of Oikonomidis [9] who utilises particle swarms to optimise a rendered model. However, this is computationally intensive, requiring a GPU implementation. Ballan [10] uses multiple cameras to reduce ambiguity and occlusion, performing motion capture of both hands slowly interacting with an object. Using Rigid Body Dynamics and several heuristic driven simulations, Melax [11] determines the best model match from a subset of approximate poses. Local minimum were then resolved though asynchronous state exploration, impacting the rate of convergence. Melax also demonstrates that rigid body dynamics, when constrained using depth, can operate similarly to iterative closest point, and is effective in traversing to the basin of local minima. Over time, model based approaches can suffer from drift due to their dependency on the previous estimations. Therefore augmentation can be used to acquire a labelling of the hand. Chua [12] uses coloured markers to aid in reducing the number of degrees of freedom, optimising model fitting, while Aristidou [13] uses optical markers and compute the remaining joints using inverse kinematics.

B. Discriminative Modelling

Discriminative based methods model the transformation from visual features in either depth or appearance to determine an unobserved prediction. This can be both in terms of classifying region labels or regressing joint positions, allowing real-time performance using a RDF. Shotton [14] used depth based features as a means of segmenting the regions of a persons body into discrete joint based regions. Keskin [2] then applied this approach for determining regions, to the hand. Keskin [15] later extended this hierarchically by specialising multiple RDFs into cluster based experts. These discriminative methods however require large amounts of labelled training data, which has previously been synthetically generated. While synthetic data can provide the vast amount of training examples required, the quality of said data is heavily dependent on both the physiological accuracy of the model and how closely the data reflects the characteristics of the capture device. To promote realism, Xu [3] incorporates the traits of shadowing and missing depth to the training images, indicative of structured light based depth. While Tang [1] explores introducing real data into training using 1200 manually labelled images. Tang acknowledges that "manually labelled realistic data is extremely costly to obtain" and so combines real and synthetic data using semi supervised learning.

C. Shape and Structure

There are also approaches that use prior understanding in the structure of the hand. Hackenberg [16] built a part based detector that searches for tube and tip structures and combines them to form finger detections. Krejov [17] also uses the structure of the hand to find geodesic extrema as an efficient means of tracking finger tips. Athitsos [18] uses probabilistic line matching and a large synthetic dataset, to learn viable edge configurations. An in-depth review of the literature on hand pose can be found in [19]

Each category of approach have a number of advantages and disadvantages: Model based optimisation has been shown to be more computational expensive than discriminative methods, however it is capable of modelling temporal coherence between successive frames. This property is desirable but can lead to drift if not corrected. Models are also capable of fitting directly to the observation, whereas discriminative approaches optimise over the data they have seen previously. The discriminative approaches aim to directly resolve the global minimum of pose. However it is extremely challenging for existing methods to generalise to unseen data due to the complexity of the hand. Vast datasets are required to cover the pose range at the expense of an increased cost

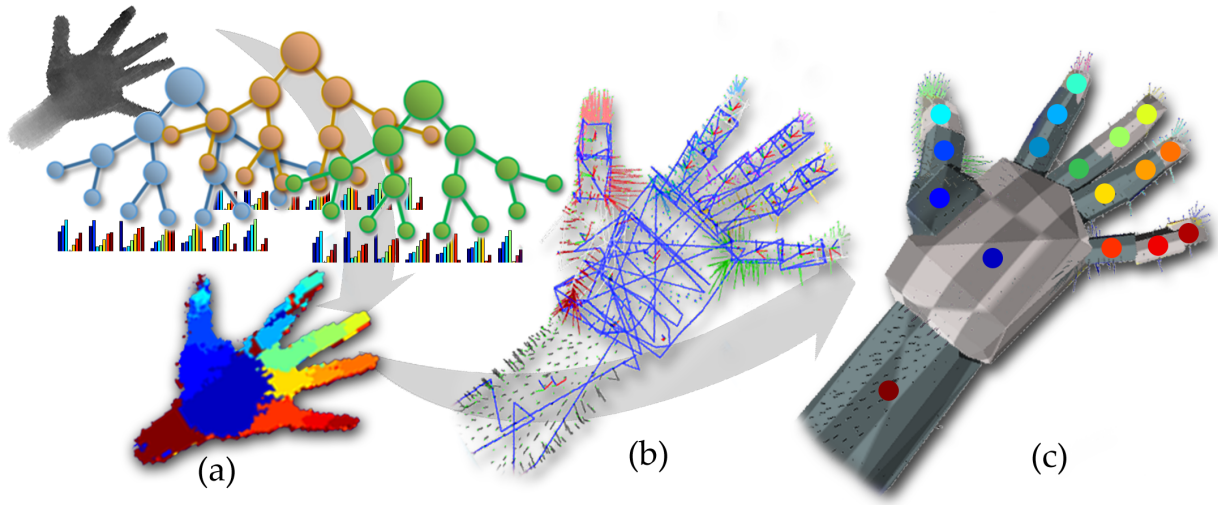


Fig. 2: System overview showing the use of an RDF to create a joint based segmentation (a). In this case the segmentation is of the full hand for clarity. This segmentation is used as a binding between the depth and the closest point on the corresponding bone (b). These constraints drive our model optimisation, resolving the final pose (c).

of computation. Approaches using shape and structure while fast, have difficulty in determining joint labelling, as fingers are visually similar.

Our proposed method uses a combination of approaches where discriminative methods approximate a global search, which initialises a model based local optimisation that descends to the basin of the global minimum. The global search is performed using a RDF, which allows for a fast coarse detection of hand regions. A local optimisation method using data driven Rigid Body Dynamics is then employed as an efficient means of modelling the hand, which implicitly handles *kinematic constraints*, prevents *self-intersection* and satisfies *temporal coherence*. These processes are also very efficient at run time and as such allow for real time performance of over 30 fps.

III. Combined descent Method

This section discusses our combined descent method which unifies discriminative and model based approaches.

Hand pose estimation can be considered a non-linear optimisation problem where the hand's appearance and model form a cost surface in parametric space. To find the hand pose, a search for the global minimum is performed. We use RDFs to provide an estimate of hand regions that guides optimisation. RDFs were first used for hand pose by Keskin [2]. Subsequent methods have shown that they are not able to fully capture the hands variation, smoothing the cost surface and inducing errors for unseen examples as can be seen in figure 3b & 3d. Our proposed method overcomes these issues through local optimisation.

If local optimisation is performed in an unconstrained fashion, with only the prior frame as an initial estimate, the model may become trapped in local minima, resulting in an incorrect estimate of pose. A guided search for the best global optimal can be performed in the model space [9] using Particle Swarm Optimisation, but this is computationally expensive. Rather than search multiple states as did Melax, our aim is to optimise only a single model, greatly reducing run time expense. We instead guide the search closer to the global optimal using the RDF.

This method consists of using labelled depth images to train a RDF to provide a sparse region segmentation, serving as an estimate of the hands configuration. This estimate binds the hand model to the observed depth using point to surface constraints to create surface based inverse kinematic relations. The constraints act similarly to springs, such that the pose from the prior frame is pulled towards the classifiers current segmentation. Rigid body dynamics are derived from Newton's laws of motion, utilising the relations between force, mass and acceleration. This means that each rigid body has velocity and acceleration attributes computed intrinsically in accordance with *temporal cohesion*. In the case of erroneous segmentation (constraints) the models prior velocity can overcome the associated incorrect force they apply. The model is refined with a fixed time step for each frame, simultaneously resolving collisions. A stable solution results that satisfies the model and the segmentation driven constraints. The final hand configuration can be represented as the centroids of each rigid body belonging to the hand model.

A. Global Estimate of Pose

To train the RDF, depth images of the hands were labelled using 17 points, each located at the centre of the bones that comprise the hand's palm and fingers. We extend the labelling to include the addition of the lower portion of the forearm. This allows for separation between the wrist and the palm of the hand. These points can be seen in figure 2c. Each depth pixel is associated with a region label, this allows the forest to partition the depth based on the regions of the hand. A three dimensional nearest neighbour assignment is used to assign each of the depth pixels to its nearest labelled

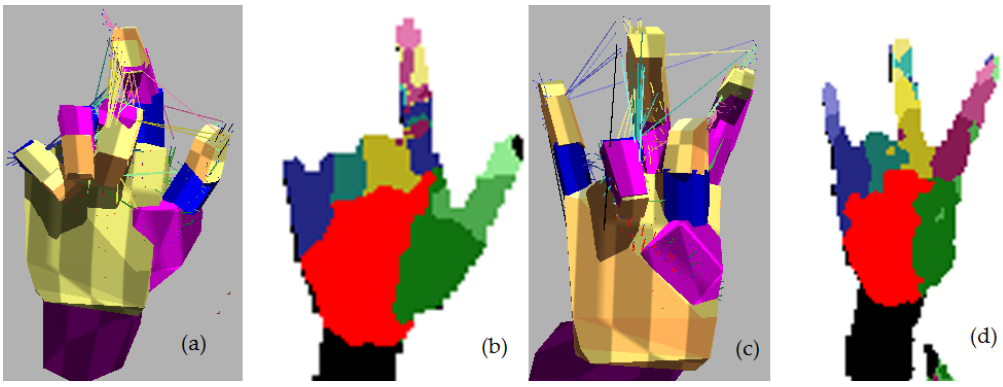


Fig. 3: The images (a) & (c) show the result of model convergence based on the classification output in the images (b) & (d) respectively. The classification of the first pose shows confusion in labelling the index finger, which without the use of our second stage refinement would be incorrectly determined. The same issue can be seen in the index and middle finger in image (d)

point to provide the region segmentation. This labelling is shown in figure 2a.

Using the nearest neighbour depth segmentation, a RDF is trained to perform pixel classification on unseen hand poses. We use a depth comparison feature which comprises of two random offset vectors $\mathbf{u}, \mathbf{v}$ whose lengths are normalised using the depth of the training sample $\mathbf{x}$, making them depth invariant. The depth at these offsets are subtracted, providing a difference in depth measure as in Equation 1.

$$F_{\mathbf{u},\mathbf{v}}(\mathbf{I}, \mathbf{x}) = \mathbf{I}\left(\mathbf{x} + \frac{\mathbf{u}}{\mathbf{I}(\mathbf{x})}\right) - \mathbf{I}\left(\mathbf{x} + \frac{\mathbf{v}}{\mathbf{I}(\mathbf{x})}\right) \quad (1)$$

where $\mathbf{I}$ is the training image.

The length of these offset vectors are also bounded by a maximal radius, allowing adjustment of how much of the hand the features can evaluate. As these features are not rotationally invariant, to account for large scale rotation, the dataset consists of rotated duplicates of the depth and segmentation. This feature was derived by Shotton [14] and has shown to be efficient to compute.

The RDF $\mathcal{F}$ is an ensemble of random decision trees, t each providing the probability of the point $\mathbf{p}$ belonging to the class l. The use of multiple trees improves classification accuracy over individual decision trees as trees within a forest optimise over varied subsets of the data, improving generalisation. During training, the root node of each random decision tree learns to partition the depth samples into left and right subsets to produce purer label distributions. The left and right subset are then propagated to subsequent nodes repeating this process until either the node has a pure distribution or a maximum depth is reached. The node learns to partition the data by selecting the feature that produces the greatest decrease in entropy. This entropy is calculated using the Shannon entropy for a random subset of features computed at training time. Each tree is trained using a fixed size subset of hand pixels from each image, to allow generalisation to the variety of poses.

During the evaluation stage, an input depth image $\mathbf{I}$ is mapped into a point cloud $\mathcal{P}$ using the camera intrinsic parameters. This point cloud is then filtered and subsampled to remove noisy outliers and to reduce the cloud to a sparse representation $\mathcal{P}_f$ that is more efficient for subsequent processing. Filtering uses a voxel grid subsampling where points in each voxel are represented using their centroid. The sub-sampled point cloud is then classified into appropriate regions provide the label l of each voxel in the hand using the RDF Eq. (2 & 3). This provides sub sampled labelling of the depth which serves as our initial estimate of pose.

$$P(l|\mathbf{I}, \mathbf{p}) = \frac{1}{|\mathcal{F}|}\sum_{t\in\mathcal{F}} P_t(l|\mathbf{I}, \mathbf{p}) \quad (2)$$

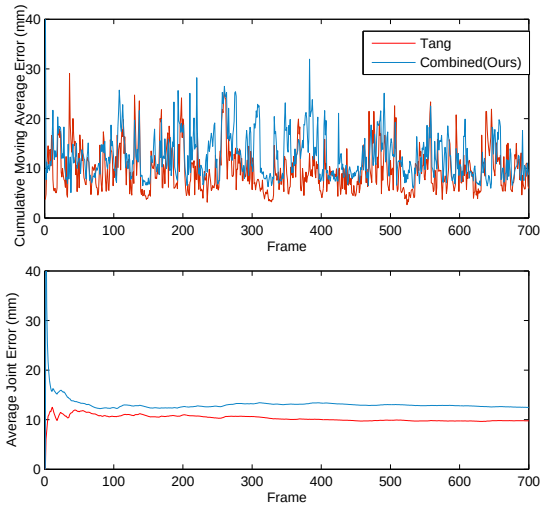
$$L(\mathbf{p}) \overset{\text{def}}{=} \arg\max_{l\in 1..n}(P(l|\mathbf{I}, \mathbf{p})) \quad (3)$$

where $L(\mathbf{p})$ is the function used to label a depth sample.

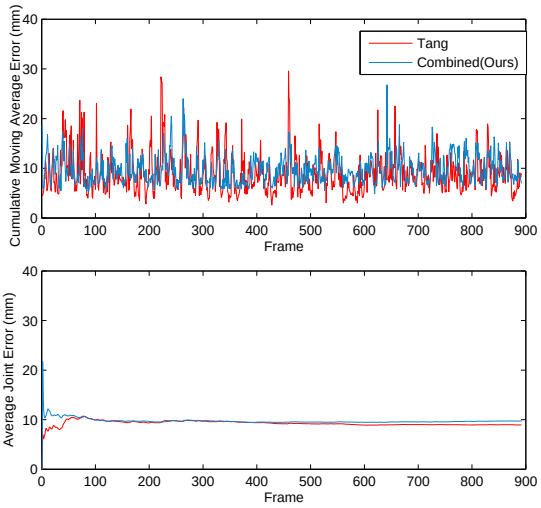
B. Hand Model Convergence

Rigid Body Dynamics simulations are a common tool in both the games and film industry [20]. The interaction of solid bodies is simulated in a 3-D environment calculating the collision between moving objects, and preventing intersection. The simulation is also capable of simulating constraints that bind objects together, for example the joints between two finger bones. We utilise a rigid body simulation to perform an efficient fitting between the hand model and classified depth.

A mean kinematic hand model $\mathcal{H}$ was constructed using proportions from several reference images of independent users, using the in-depth study of Segmentos [21]. The hand model is comprised of $n = 17$ convex rigid bodies $\mathcal{H} = b_1 \cup b_2 \cup ... \, b_n$ (which includes the lower portion of the forearm). These convex shapes (bones) have a low polygon count to allow fast computation of the subsequent steps while still being representative of the hand. The bones are connected using rotational constraints which have a limited range of rotation to reflect the kinematic limitations of the hand [22]. The mass of each component was estimated using the size of each convex shape.

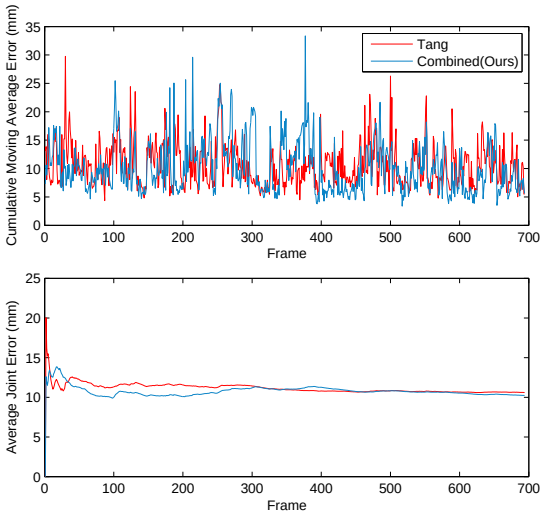


(a) Sequence one using autonomous ground truth

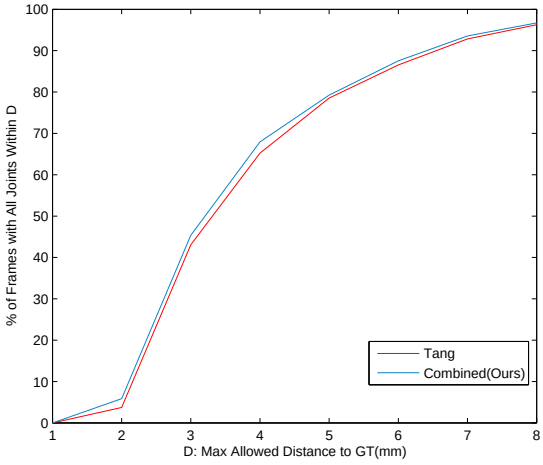


(b) Sequence two using autonomous ground truth

Fig. 4: Evaluation using autonomously labelled ground truth comparing the per frame mean joint error, and it's cumulative moving average



(a) Performance against the manually corrected ground truth of sequence one



(b) Worst-case accuracy of the manually labelled sequence also showing an improvement over state of the art

Fig. 5: Evaluation using corrected ground truth from sequence 1 showing improved performance.

The binding between the depth and model is achieved using point to surface constraints. Each constraint comprises of a sample point $\mathbf{p} \in \mathcal{P}_f$ taken from the filtered point cloud, and corresponding rigid body b_i which is determined by the RDF's classification output. These two components form constraints for the closest point on the surface $\mathbf{p}'$ (Eq. 4).

$$\mathbf{p}' = \underset{\mathbf{p}_b \in b_i}{\arg\min}(||\mathbf{p} - \mathbf{p}_b||), \text{where } i = L(\mathbf{p}) \tag{4}$$

This closest point is found using a derived Gilbert-Johnson-Keerthi (GJK) [23] distance algorithm. GJK is an efficient means of finding the closest points between two polygonal objects.

Once the hand is bound using constraints to the depth, the simulation is iterated for a fixed duration allowing the model to converge on the depth appearance. This simulation comprises of a broad and narrow pass for handling object collisions, which prevents the fingers from intersecting, satisfying the precondition of no *self-intersection* using a constraint solver. The constraint solver optimises the impulse forces to satisfy the point to surface constraints. This completes one iteration of the simulation. The constraints are then updated to reflect the new closest surface point $\mathbf{p}_b$ and the process repeated until convergence.

As the constraints bind the hands surface to the depth, the model converges to reside within the hand's depth contour, satisfying the *depth observation*.

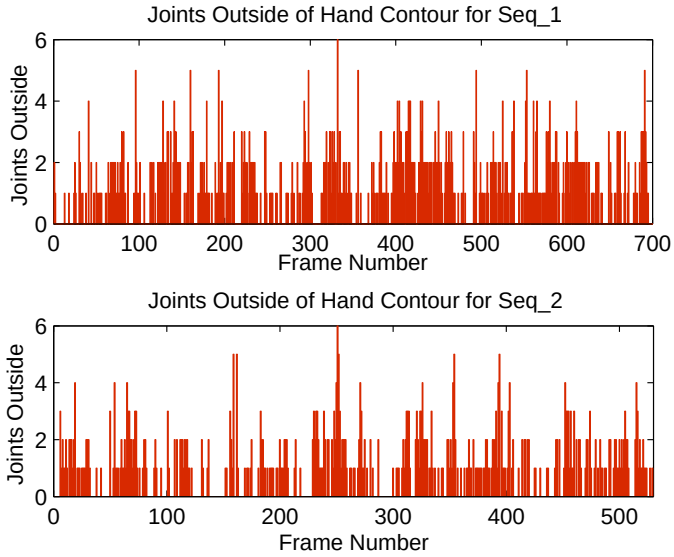
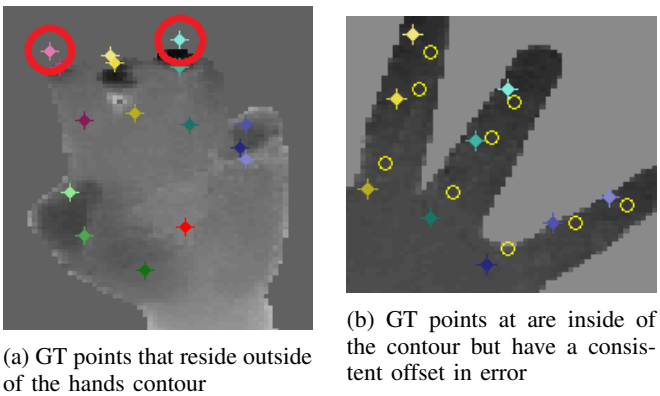


Fig. 6: Graph showing labelling error in ground truth of test data. Measured using joints outside of the hand contour in seq_1 and seq_2. This error is due to the test sequence having been labelled using automatic means.



(a) GT points that reside outside of the hands contour

(b) GT points at are inside of the contour but have a consistent offset in error

Fig. 7: The coloured markers show the ground truth provided with the sequence. Yellow circles represent the result from our combined approach.

IV. Evaluation

A. Comparison with state of the art

In the following experiments, we use an RDF that was trained using optimised parameters against a validation set with the objective of classification accuracy. The resulting forest of 3 trees was trained using 20,000 images to a depth of 20, using 2000 randomised offsets and a offset radius of 20 pixels. 100 threshold selections were made at each node and samples were acquired using a random sampling of 1000 pixels per image.

B. Evaluation using Autonomously Labelled data

We compare our result using the test sequences and resulting joint positions provided by Tang [1]. Tang has shown a significant improvement in accuracy over the use of RDFs alone [2] and Real Time Model optimisation [11]. Each sequence consists of approximately 700 frames that have been automatically labelled using the approach of Melax [11] and manually corrected in failure cases. We found that there are large inaccuracies in the resulting labelling which can be expected as obtaining a ground truth of high accuracy is very challenging. In figures 4a & 4b we show our performance against that of Tangs using the mean error between the joints and their corresponding ground truth, on a per frame basis. We also demonstrate performance using the cumulative moving average of this error, which is the normalised mean error of all previous frames (lower is better)

When using the erroneous automatically obtained ground truth. It can been seen that there is a persistent error between our performance and theirs. When looking at the failure cases of the autonomously generated labels 7a it can be seen that a number of joints reside outside of the contour of the hand which constitutes as an invalid joint position. We also demonstrate this in figure 6 where we quantify this with a naive measure by calculating the number of joints that are positioned outside of hand's contour on a per frame basis. There is also error observed for labelled points inside the contour as shown in figure 7a Therefore, to perform a fair evaluation, we corrected the ground truth of sequence 1 with manual annotation which we make available at personal.ee.surrey.ac.uk/Personal/P.Krejov/.

C. Evaluation using Manually Corrected data

When using the manually annotated ground truth, figure 5a shows that our approach provides a more accurate joint localisation. This improvement is due to our approach converging successfully with the depth and is confirmed in the examples shown in figure 8. We also compute the worst case accuracy in joint estimation finding that 68% of joints are within 40 mm of the ground truth. This is a 3% increase over state-of-the-art as seen in figure 5b.

These results show that the combined method can accurately determine the joint positions within the centre of the finger. The direct regression method proposed by Tang however, exhibits the noise characteristics associated with the automatic labelling used in training.

Convergence rate also shows significant improvement when compared against that of Melax's approach. Melax demonstrates model convergence for fast moving poses as taking 15 to 30 frames, while the combined method converges within 3 to 5 frames.

V. Conclusions

In this paper we presented a method for real-time hand pose estimation. The approach utilised a RDF to provide a initial estimate, which is then employed in a local optimisation strategy with the depth using a Rigid Body Dynamics simulation. This combines the benefits of both approaches to perform successful hand pose estimation and demonstrates state-of-the-art performance. As future work, we aim to investigate fine-grained adaptation of the hand model to specific users over the course of interaction.

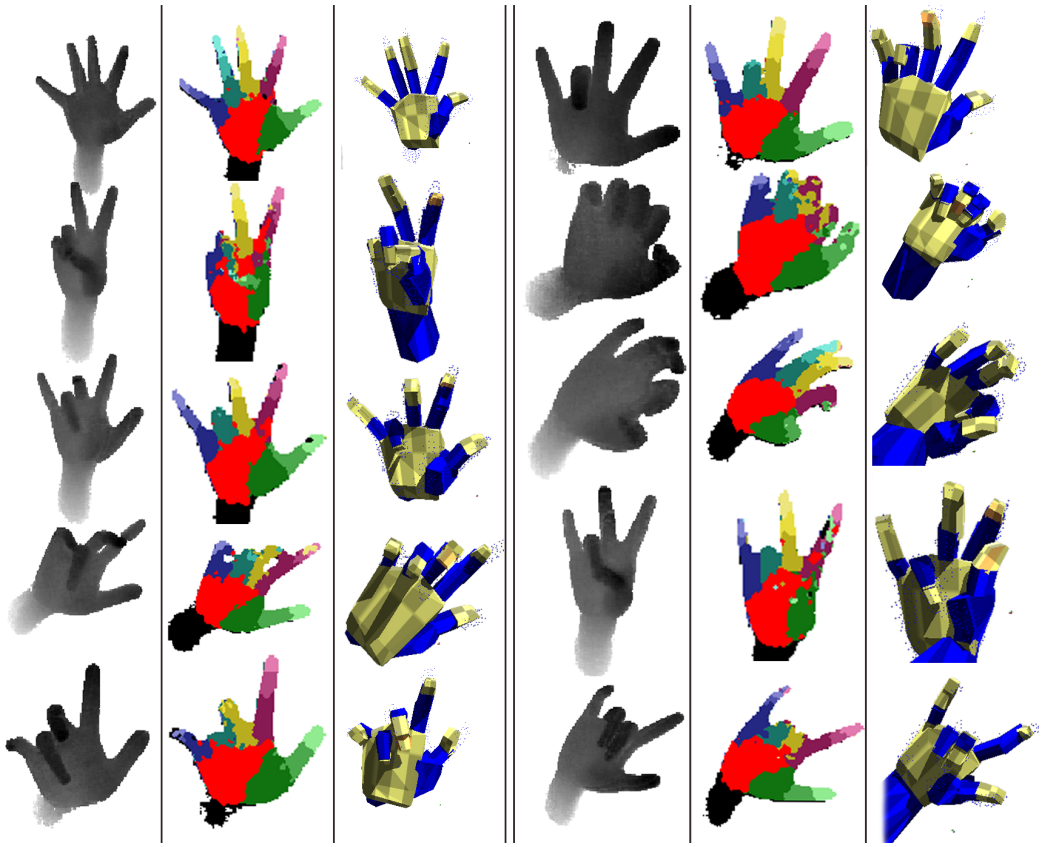


Fig. 8: This figure show qualitative results the combined method. Showing the raw depth, forest segmentation and converged hand model.

REFERENCES

[1] Tang, D., Yu, T., Kim, T.: Real-time Articulated Hand Pose Estimation using Semi-supervised Transductive Regression Forests. In: ICCV. (2013)

[2] Keskin, C., Kirac, F., Kara, Y.E., Akarun, L.: Real time hand pose estimation using depth sensors. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), IEEE (November 2011) 1228–1234

[3] Xu, C., Cheng, L.: Efficient Hand Pose Estimation from a Single Depth Image. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 3456–3462

[4] Feng, L., Po, L., Xu, X.: An adaptive background biased depth map hole-filling method for Kinect. ...Society, IECON 2013 ...(2013) 2366–2371

[5] Sudderth, E., Mandel, M., Freeman, W., a.S. Willsky: Visual Hand Tracking Using Nonparametric Belief Propagation. 2004 Conference on Computer Vision and Pattern Recognition Workshop (2004) 189–189

[6] Stenger, B., Thayananthan, A., Torr, P.H.S., Cipolla, R.: Model-based hand tracking using a hierarchical Bayesian filter. IEEE transactions on pattern analysis and machine intelligence **28**(9) (September 2006) 1372–84

[7] Hamer, H., Schindler, K.: Tracking a hand manipulating an object. ...Vision, 2009 IEEE ...(2009)

[8] de La Gorce, M., Fleet, D.J., Paragios, N.: Model-Based 3D Hand Pose Estimation from Monocular Video. IEEE transactions on pattern analysis and machine intelligence (February 2011) 1–15

[9] Oikonomidis, I., Kyriazis, N., Argyros, A.: Efficient model-based 3D tracking of hand articulations using Kinect. BMVC (2011) 1–11

[10] Ballan, L., Taneja, A.: Motion Capture of Hands in Action using Discriminative Salient Points

[11] Melax, S., Keselman, L., Orsten, S.: Dynamics based 3D skeletal hand tracking. In: Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games - I3D '13, New York, New York, USA, ACM Press (2013) 184

[12] Chua, C.S., Guan, H., Ho, Y.K.: Model-based 3D hand posture estimation from a single 2D image. Image and Vision Computing **20**(3) (March 2002) 191–202

[13] Aristidou, A., Member, S., Lasenby, J.: Motion Capture with Constrained Inverse Kinematics for Real-Time Hand Tracking. (March) (2010) 3–5

[14] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR 2011, IEEE (June 2011) 1297–1304

[15] Keskin, C., Kraç, F., Kara, Y., Akarun, L.: Hand pose estimation and hand shape classification using multi-layered randomized decision forests. Computer VisionECCV 2012 (2012) 852–863

[16] Hackenberg, G.: Lightweight Palm and Finger Tracking for Real-Time 3D Gesture Control. (March 2010) (2011) 19–26

[17] Krejov, P., Bowden, R.: Multi-touchless: Real-time fingertip detection and tracking using geodesic maxima. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE (April 2013) 1–7

[18] Athitsos, V., Sclaroff, S., Street, C.: Estimating 3D Hand Pose from a Cluttered Image. (2003)

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

[20] Bulletphysics.org: Real-time physics simulation (2014)

[21] Segmentos, P.D.L., Mano, D.: Proportions of Hand Segments. **28**(3) (2010) 755–758

[22] Albrecht, I., Haber, J., Seidel, H.p.: Construction and Animation of Anatomically Based Human Hand Models. (2003)

[23] Gilbert, E., Johnson, D., Keerthi, S.: A fast procedure for computing the distance between complex objects in three-dimensional space. IEEE Journal on Robotics and Automation **4**(2) (April 1988) 193–203