Unified 2D and 3D Hand Pose Estimation from a Single Visible or X-ray Image

Akila Pemasiri
akila@cse.mrt.ac.lk
Kien Nguyen
k.nguyenthanh@qut.edu.au
Sridha Sridharan
s.sridharan@qut.edu.au
Clinton Fookes
c.fookes@qut.edu.au

Image and Video Research Lab
Queensland University of Technology
2 George Street
GPO Box 2434, Brisbane QLD 4001, Australia

Abstract

Robust detection of the keypoints of the human hand from a single 2D image is a crucial step in many applications including medical image processing, where X-ray images play a vital role. In this paper, we address the challenging problem of 2D and 3D hand pose estimation from a single hand image, where the image can be either in the visible spectrum or an X-ray. In contrast to the state-of-the-art methods, which are for hand pose estimation on visible images, in this work, we do not incorporate the depth images to the training model, there by making the pose estimation more appealing for the situations where the access to the depth images is not viable. Besides, by training a unified model for both X-ray and visible images, where each modality captures different information which complements each other, we elevate the accuracy of the overall model. We present a cascaded network architecture which utilizes a template mesh to estimate the deformations in the 2D images where the estimation is propagated in different cascaded levels to increase the accuracy.

1 Introduction

Hand pose estimation is a long-standing challenge in computer vision and image processing research fields. There are numerous applications of hand pose estimation including human-computer interaction, virtual reality [23, 31], robotics [8], and medical image analysis [7]. In medical image analysis, 2D pose estimation plays a vital role in tasks such as organ motion correction [6] and deformable image registration [4].

Hand pose estimation of visible images is widely incorporated in many clinical applications where non-invasive monitoring of patients is performed, including those who have motor disorders as well as in tremor diagnosis [15, 35, 42]. In assessing skeletal maturity of a patient's hand using X-ray images, hand pose estimation is used as an essential subroutine [10, 25]. The main features that are utilized in this process are the morphological patterns at the bone joints and bone joint identification remains a crucial process in automated radiology [30].

Moreover, the semantic keypoints resulting from 2D/3D pose estimation are widely used in 3D model reconstruction processes, where the detected keypoints are used as anchors [1, 32]. In the visible image domain, reconstructing the 3D model from a single RGB image is of broad and current interest [9, 37, 38]. The importance of 3D model reconstruction is not restricted to the visible images; significant attention has been devoted to the 3D reconstruction of images in other modalities as well [5, 22, 24, 29]. Specifically, in the medical imaging domain, reconstruction of 3D models from X-ray images has been very useful in the pathological analysis [11, 14]. Keypoint identification on the X-ray images (i.e., based on fiducial markers or automatic) is a fundamental step in 3D reconstruction from X-ray images [2, 16].

Analysing the images of an object in different modalities, where the presented details complement each other is of great importance in many computer vision tasks including medical [18, 27] applications. In addition to the complementary information, one of the main advantages in using different imaging modalities is that the modalities which can be captured abundantly with minimum cost and with minimum effort (e.g., RGB images), can be used in combination with other imaging modalities which require specific environments, costly equipment and which may also be associated with harmful radiations (e.g., X-ray and CT images) [34]. For an example current methods of 3D reconstruction from X-ray images require more than a single image; and using an X-ray image in combination with visible images can eliminate the need for multiple X-ray images. However, this necessitates a keypoint identification framework which can accurately identify keypoints of both X-ray and visible images.

To address the above issues we present in this paper, a unified method for 2D/3D hand pose estimation for visible and X-ray images. To our best knowledge this is the first effort which targets 3D hand pose estimation for X-ray images. The main contributions of this paper are summarised below:

- In contrast to the existing approaches which consider only an individual modality, our framework considers both the bone structure of the human hand which enables the estimation of deformation in the X-ray and as the muscle structure along with the skin, which contributes to the visible image deformation estimation.

- In contrast to the prior work, which uses depth images in developing models for hand pose estimation from an RGB image, we do not use the depth images; as such we increase the potential of using our model for the datasets which do not have the corresponding depth images.

- The existing hand pose estimation methods for X-ray images [12, 25] have targeted only the canonical correspondence of images. In contrast, our framework looks on semantic correspondence, which allows greater flexibility by identifying the corresponding points in the images where there are drastic deformations.

- We provide manual keypoint annotation for hand X-ray images in the **mu**sculoskeletal **ra**diographs (MURA) dataset and a synthesized set of images, which replicates the structure of X-ray images with their 2D and 3D groundtruth annotation.

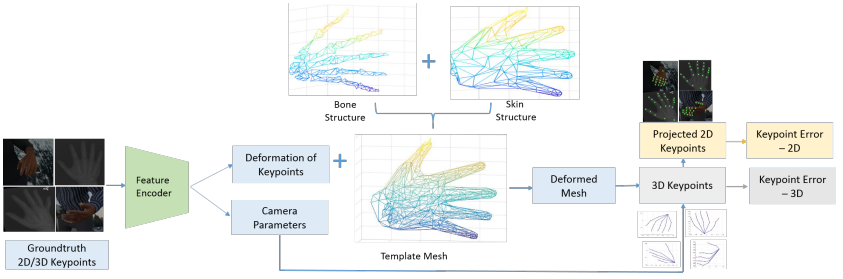


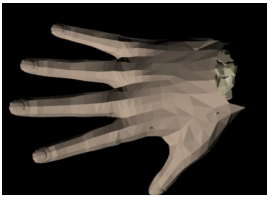
Figure 1: Overview of the proposed framework, where at the training time images from both the modalities (to reflect the bone structure and the skin structure) are fed into the network along with the the available groundtruth values (2D keypoints and 3D keypoints when available). As the feature encoder, any feature extraction framework (e.g. variations of Resnet) can be used.

2 Methodology

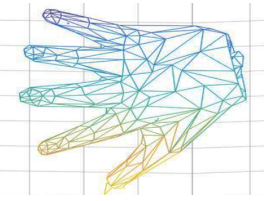
The overview of the proposed framework is depicted in Figure 1, where we use X-ray images and the visible images as the input to the framework. The groundtruth annotation of the 2D pose and 3D pose when available, are used at the training time. Based on the features obtained by encoding the input images, we obtain the camera parameters (i.e. rotation, translation and scale) and the estimated deformation of the keypoints. The deformed keypoints of the template 2D mesh are then projected using the camera model and the loss value between these projected keypoints and the groundtruth keypoints are used to measure the accuracy values. When the 3D groundtruth points are available, we estimate the 3D keypoint error as well. Each component of our framework is described in detail in the following sections of this paper.

2.1 Hand mesh representation

As the template mesh we use a 3D mesh with both bone structure and the skin structure, which are extracted using a 3D human hand model [41]. Though the model can generate meshes with high resolution, to reduce the complexity of the deep learning model we use low resolution template models; $M_{bone} \in R^{N\times 3}$ and $M_{skin} \in R^{N\times 3}$, where N denotes the number of vertices in each for the bone mesh and the skin mesh. The rendered view of the template mesh and the mesh representation for both the meshes are depicted in Figure 2. However, as a higher resolution mesh can yield better accuracy values, in our framework we considered increasing the resolution of the mesh in a cascaded manner while allowing the deep learning model to maintain the simplicity among each of the levels in the cascade. MANO [36] is another hand model representation method, which is parameterized by shape ($\overrightarrow{\beta} \in R^{10}$) and pose values ($\overrightarrow{\alpha} \in R^{k\times 3}$), where k is the number of keypoints. However, generating this model requires a large number of 3D hand scans. Therefore it is not feasible to use such a model for X-ray pose estimation as obtaining a very large number of hand scans with bones only is not viable.

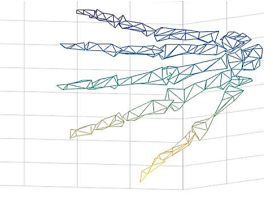


(a) Skin model of the hand



(b) Mesh of the skin model

(c) Bone structure of the hand



(d) Mesh of the bone structure

Figure 2: The individual meshes that were used as the template mesh representation. These were combined together to incorporate the details of bone structure and the skin structure.

2.2 Framework architecture

The details of the framework architecture is described in the following sections.

Feature encoding The detailed network architecture of our framework is depicted in Figure 3, where the Resnet-18 [13] is used as the feature extraction framework. Alternatively, other network architecture variants including Resnet-50, Resnet-101 and Resnet-152 can be used in this framework. The framework, illustrated in Figure 3 uses the cascaded architecture where the deformation of keypoints is performed at several levels. Note that, depending on the required accuracy levels, the number of deformation blocks and the Resnet convolution layer number of which the features are extracted from can be adjusted.

Camera parameters and keypoint deformation estimation The output of each feature extraction layer is passed through another encoder layer, which has multiple fully connected layers, which are followed by several network segments where each will estimate a different output. These network branches include (1) camera estimation networks; which are "Rotation Predictor", "Scale Predictor", and "Translation Predictor" (2) "Deformation Estimator" which estimates the keypoint deformation. Each of these segments is constituted of fully connected layers where the rotation predictor, translation predictor and scale predictor have 4, 2 and 1 output units respectively and the deformation predictor has $N \times 3$ outputs, where N is the number of vertices in the considered template mesh. The deformation values associated with the keypoint vertices are added to the template mesh to get the deformed keypoint location, and in the setting of cascaded network architecture, the deformation of all the vertices are added to the template mesh to generate the new template mesh for the next cascade level.

Cascaded architecture As mentioned earlier the cascaded architecture is used to accommodate the balance between complexity of the model and the required accuracy levels. The use of graph-unpooling [42] in the cascaded architecture will increase the resolution of the mesh, which leads to the refined keypoint identification. Since edge-based unpooling yields a uniform upsampling on the mesh [42], in this framework we applied edge based unpooling

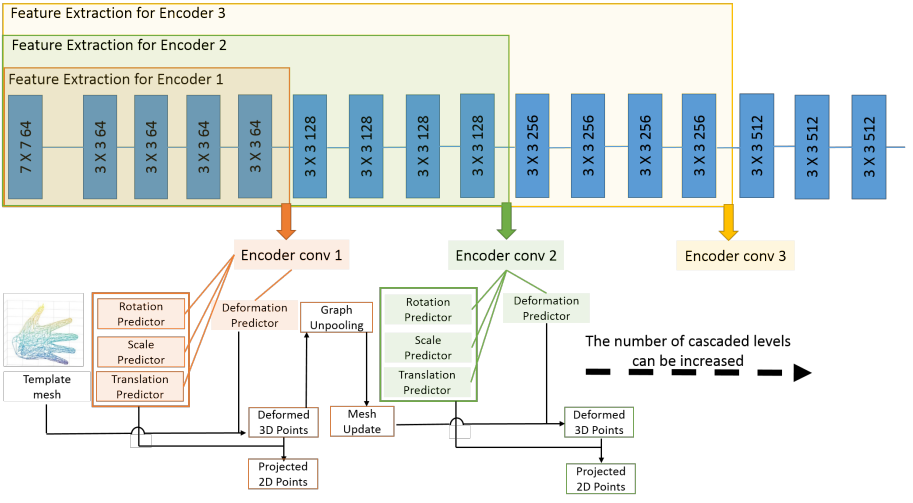


Figure 3: The detailed network architecture which uses Resnet-18. This architecture shows 2 cascade levels, where the output mesh of the 1st level is enhanced using graph unpooling.

to the meshes under consideration.

Vertices to keypoint association Associating the vertices with the keypoints is another major consideration in our framework. Associating the vertices manually with the keypoints will reduce the adaptability of our framework, as a change in the initial mesh will lead to changes in all the steps. Instead, we use a keypoint assignment matrix $A_{k\times N}$ where k is the number of keypoints and N is the number of vertices in the corresponding template mesh, which will learn the association between the vertices and the keypoints [19].

Learning We design a loss function to aggregate four losses incurred by the estimation accuracy of 2D and 3D keypoints, where

$$\begin{aligned} Loss = w_1 \left(\alpha_1 L_{kp\_2D} + \beta_1 L_{kp\_3D} + \gamma \delta_1 L_{v2kp\_visible} + (1-\gamma)\,\delta_1 L_{v2kp\_Xray}\right) \\ + w_2 \left(\alpha_2 L_{kp\_2D} + \beta_2 L_{kp\_3D} + \gamma \delta_2 L_{v2kp\_visible} + (1-\gamma)\,\delta_2 L_{v2kp\_Xray}\right). \end{aligned} \tag{1}$$

The loss function denoted in Equation 1, corresponds to a network architecture that has 2 cascaded levels. The hyper parameters of our loss function includes w_1, w_2, which assigns the weights for the each cascade level, and α_i, β_i and δi indicate the weights assigned for the 2D keypoint loss, 3D keypoint loss and the keypoint-vertex association loss respectively. In Equation 1, γ is set to 1 if the input image is a visible image. We have reflected the effect of input image modality only on the keypoint-vertex association loss, as the keypoint-vertex association error should converge in a modality specific way; i.e.: the keypoints associated with X-ray images should be reflected on the bone structure and the keypoints associated with visible images should be reflected on the skin structure. For the keypoint error we used mean squared error between the groundtruth keypoints and the estimated keypoints (Equation 2), where each 2D keypoint is defined as $[x_p, y_p]$ and each 3D keypoint is defined as $[x_p, y_p, z_p]$. The keypoint-vertex association loss we used is the average cross entropy loss (Equation 3) for each keypoint in the total vertex distribution. In equation 3, A denotes the keypoint assignment matrix and A_{ij} denotes the element in the i^{th} row and the j^{th} column of

that matrix .

$$L_{kp} = \frac{1}{k}\sum_{i=1}^{k}\left\| kp_{gt}^{(i)} - kp_{pred}^{(i)} \right\|_2^2 \tag{2}$$

$$L_{v2kp} = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{N} A_{ij} log(A_{ij}) \tag{3}$$

3 Experiments

In this section we describe the datasets that we used, the ablative study we have conducted to evaluate the effectiveness of each of the components in our architecture, and the experimental results that we obtained in benchmarking our method with the state-of-the-art.

3.1 Datasets

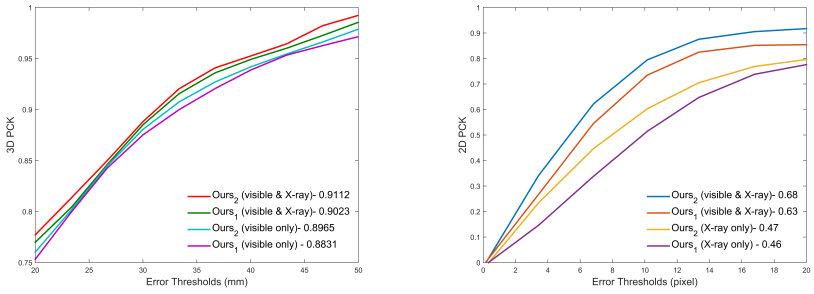
This work utilizes two main types of datasets 1) Visible image dataset and 2) X-ray image dataset. Though there are publicly used datasets for hand keypoint estimation for visible images, there is no publicly available dataset for hand pose estimation of the X-ray images.

Visible hand pose datasets: In this evaluation we use Stereo Handpose Dataset (STB) [46], Rendered Hand Pose Dataset (RHD) [48] and Dexter-Object Dataset [40]. Following the evaluation protocol used in the most related literature [26, 43, 45], for the STB dataset evaluation we used 10 image sequences for training and 2 image sequences for testing and for the RHD dataset we used its training and testing tests respectively. Dexter dataset was used to evaluate the generalization capability of our models which were trained on other datasets.

X-ray hand pose datasets: For the X-ray hands, the main challenge we encountered was that of not having a large dataset with 2D and 3D pose estimation. From the MURA (musculoskeletal radiographs) dataset which contains X-ray images of [33] human upper extremity, we used the hand X-ray images and manually annotated 1000 hand images, with each hand having 21 keypoints, and then by augmenting that data we increased the X-ray hand dataset with 2D annotation. To yield better accuracy for 3D pose estimation on X-ray images, it is mandatory to have X-ray images with both 2D and 3D annotations. To accomplish this we used a 3D model of the human anatomy where the texture of the model was used to mimic properties of X-ray images. The rendered images were considered as the corresponding 2D X-ray images, and were subjected to 2D keypoint annotation and from the 3D-coordinates of the model we obtained the 3D coordinates of points that are associated with the keypoints.

3.2 Evaluation Matrix

For our ablative studies and for the benchmarking with the state-of-the-art methods we used Percentage of Correct Keypoints (PCK) score, which is widely used as the pose estimation accuracy measurement [5, 17, 26, 28, 39, 47]. When estimating the PCK, if the predicted keypoint lies within a circle (for 2D pose) or within a sphere (for 3D pose) with a given radius with respect to the groundtruth value, it is considered as a correct keypoint.



(a) 3D PCK on RHD visible dataset (b) 2D PCK on MURA X-ray dataset

Figure 4: **Ablation study results:** (a) 3D PCK values we obtained for RHD visible dataset and Figure (b) 2D PCK values we obtained for MURA X-ray image dataset.

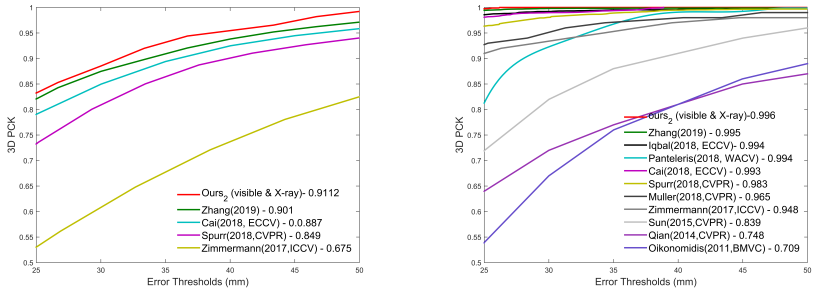
3.3 Ablative Studies

In our framework, we used Resnet-18 architecture as the feature extraction framework, with two different configurations where cascade level 1 and cascade level 2 are used. In recording our results, we have denoted these two configurations as $ours_1$ and $ours_2$. We trained our network on three different dataset configurations; 1) Visible images only, 2) X-ray images only and for 3) Visible and X-ray images jointly. All the trained models were trained using the Adam optimizer [21], with a learning rate of 0.001 and with a momentum of 0.9. In $ours_2$ configuration $w1$ and $w2$ were set to 0.5 and α_i, β_i and δ_i were set to 0.4, 0.4 and 0.2 respectively.

For the RHD dataset, the obtained 3D PCK values are depicted in Figure 4(a), where the cascade levels have positively affected the accuracy of overall results. Furthermore, it can be observed that the model trained with both visible and X-ray images have yielded better accuracy values for both network architectures, making it evident that the complementary information from different modalities can enhance the performance. The same observation was encountered for the MURA X-ray image dataset (Figure 4 (b)), and it should be noted that compared to the visible image dataset size (41,258 images), the X-ray image dataset used was small (6,000 images). Getting better accuracy levels for X-ray 2D pose estimation, when the model was trained on X-ray and visible image dataset demonstrates the modalities where the samples are abundant (e.g. visible) can be effectively utilized to enhance the performance on other imaging modalities, where the samples are scarce (e.g.X-ray images).

3.4 Comparison to the state-of-the-art

3D pose estimation on visible images: Figure 5 depicts the 3D PCK values that we obtained for our best configurations , compared to the 3D PCK values that have been obtained in state-of-the-art methods [26, 28, 47, 48]. To make our results comparable to the previous approaches, we used the same evaluation protocol that has been followed in the literature [26, 47, 48]. It can be observed that when considering STB dataset, which is extracted from videos and hence lacks diversity, our method yields comparable results with the state-of-the-art. For the RHD dataset which contains images from a wide range of backgrounds, subjects and poses, our method has outperformed the state-of-the-art method by a considerable margin.



(a) 3D PCK on RHD dataset (b) 3D PCK on STB dataset

Figure 5: PCK values based on 3D pose estimation on visible images compared with the state-of-the-art (a) RHD dataset and (b) STB dataset.

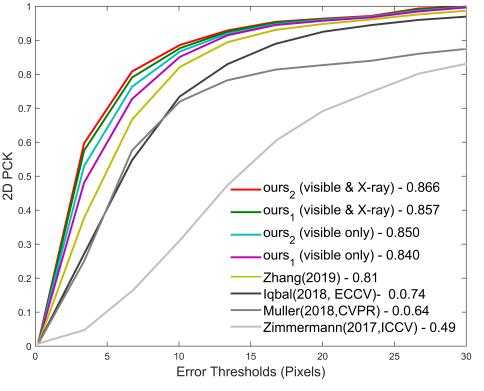


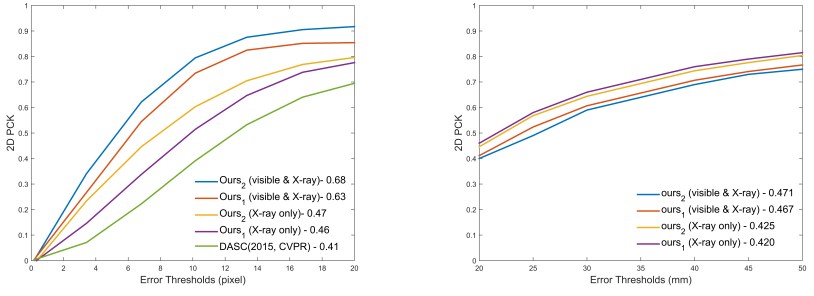
Figure 6: PCK values based on 2D pose estimation on visible images (Dexter-Object dataset) compared with the state-of-the-art. Similar to the existing benchmark methods we trained the model on RHD and STB datasets and tested on Dexter-Object dataset ($ours_1$(visible only) and ($ours_2$(visible only) indicate the results). Then to assess the impact of using visible and X-ray images jointly, we trained our model on RHD, STB and X-ray datasets and tested on Dexter-Object dataset ($ours_1$(visible & X-ray) and ($ours_2$(visible & X-ray) indicate the results).

2D pose estimation on visible images: Following the common evaluation protocol, we experimented our models on Dexter-Object dataset for the generalizing capability of the model. We trained the models on RHB, STB and X-ray dataset and tested on Dexter-Object dataset. The obtained 2D PCK values for the Dexter-Object dataset are depicted in Figure 6, and it can be observed that our method under all configurations has outperformed the state-of-the-art [17, 26, 47, 48]. It should be noted that the results recorded for exsiting methods in Figure 5 and 6 are extracted from the respective publications by the original authors. Furthermore, the observations related to multi-modal dataset training that were made in the ablation study are reflected here as well.

2D pose estimation on X-ray images: For the X-ray images, to benchmark our method we used Dense adaptive self-correlation descriptor for multi-modal and multi-spectral cor-

respondence (DASC) [20], which aims to identify the correspondence between multimodal images. In the benchmarking process, for the DASC we used a specific X-ray image as the source image and by changing the target images, we identified the corresponding points. The 2D PCK values that we obtained by using DASC and our method are depicted in Figure 7 (a). We used the annotated MURA dataset for the evaluations.

3D pose estimation on X-ray images: To the best of our knowledge, this is the first paper to estimate the 3D pose in X-ray images and the PCK values associated with the 3D X-ray pose estimation is depicted in Figure 7 (b). The impact on using joinltly using visible images with X-ray images can be clearly observed in the increased PCK values. It is important to note that this model has been trained only on a small dataset of synthesized X-ray images



(a) 2D PCK on X-ray images (b) 3D PCK on X-ray images

Figure 7: PCK values based on 2D and 3D pose estimation on X-ray images: (a) Mura X-ray dataset and Figure (b) Synthesized X-ray image dataset.

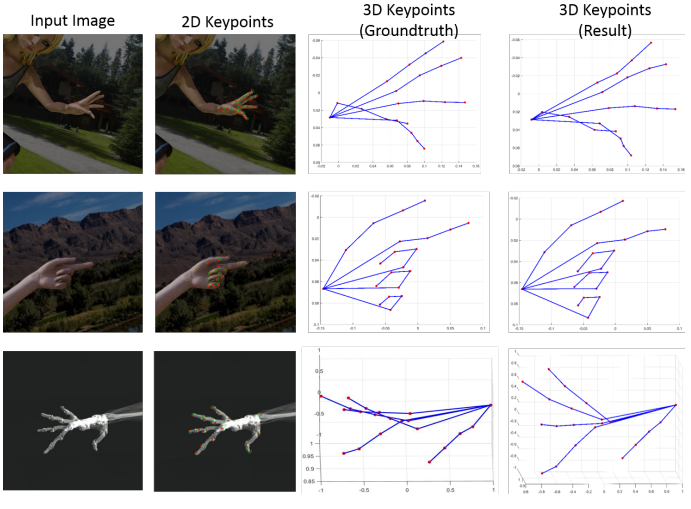


Figure 8: **Qualitative examples:** The first column shows the input image, and in the second column the groundtruth keypoints are marked in green circles and the keypoints obtained by our framework are marked in red asterisks. The third column illustrates the groundtruth 3D keypoints and the fourth column illustrates the 3D keypoints obtained from our framework. More qualitative results can be found in the supplementary material.

with 3D annotations. Figure 8 depicts some of the qualitative results that we obtained from our framework.

4 Conclusion

While hand pose estimation has been evolved through the decades, there is no exploration of how multi-modal images can be used to enhance the performance of hand pose estimation. Furthermore even though the hand key point estimation on X-ray images has been a subject of research for a considerable time, the 3D pose estimation from a single X-ray image has not been researched. In this paper, we have presented a modality invariant 2D/3D hand pose estimation method for visible and X-ray images. Using a single model for both the imaging modalities is a key innovation in our approach which enables the model to capture features which complement each other, resulting in improved accuracy values. Note also that our approach enables efficient use of data in image modalities where the samples are abundant (i.e. visible images) to enhance the performance of modalities where the data is sparse (i.e. X-ray images). Through extensive evaluations, we confirm that our framework outperforms the state-of-the-art methods for 2D and 3D hand pose. Our work is also the first effort in 3D hand pose estimation from a single X-ray image.

References

[1] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M Seitz, and Richard Szeliski. Building rome in a day. In *2009 IEEE 12th international conference on computer vision*, pages 72–79. IEEE, 2009.

[2] Aurelien Baudoin, Wafa Skalli, Jacques A de Guise, and David Mitton. Parametric subject-specific model for in vivo 3d reconstruction using bi-planar x-rays: application to the upper femoral extremity. *Medical & biological engineering & computing*, 46(8): 799–805, 2008.

[3] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 666–682, 2018.

[4] Edward Castillo, Richard Castillo, Josue Martinez, Maithili Shenoy, and Thomas Guerrero. Four-dimensional deformable image registration using trajectory modeling. *Physics in Medicine & Biology*, 55(1):305, 2009.

[5] Chia-Yen Chen, Chia-Hung Yeh, Bao Rong Chang, and Jun-Ming Pan. 3d reconstruction from ir thermal images and reprojective evaluations. *Mathematical Problems in Engineering*, 2015, 2015.

[6] Haili Chui and Anand Rangarajan. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, 89(2-3):114–141, 2003.

[7] Lazzaro di Biase, John-Stuart Brittain, Syed Ahmar Shah, David J Pedrosa, Hayriye Cagnan, Alexandre Mathy, Chiung Chu Chen, Juan Francisco Martín-Rodríguez, Pablo Mir, Lars Timmerman, et al. Tremor stability index: a new tool for differential diagnosis in tremor syndromes. *Brain*, 140(7):1977–1986, 2017.

[8] 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.

[9] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017.

[10] Arkadiusz Gertych, Aifeng Zhang, James Sayre, Sylwia Pospiech-Kurkowska, and HK Huang. Bone age assessment of children using a digital hand atlas. *Computerized Medical Imaging and Graphics*, 31(4-5):322–331, 2007.

[11] Baishali Goswami and Santanu Kr Misra. 3d modeling of x-ray images: A review. *International Journal of Computer Applications*, 975:8887, 2015.

[12] Yanrong Guo, Guorong Wu, Jianguo Jiang, and Dinggang Shen. Robust anatomical correspondence detection by hierarchical sparse graph matching. *IEEE transactions on medical imaging*, 32(2):268–277, 2013.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] S Hosseinian and H Arefi. 3d reconstruction from multi-view medical x-ray images–review and evaluation of existing methods. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 40, 2015.

[15] Murtadha D Hssayeni, Michelle A Burack, and Behnaz Ghoraani. Automatic assessment of medication states of patients with parkinson's disease using wearable sensors. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6082–6085. IEEE, 2016.

[16] Ludovic Humbert, Jacques A De Guise, Benjamin Aubert, Benoît Godbout, and Wafa Skalli. 3d reconstruction of the spine from biplanar x-rays using parametric models based on transversal and longitudinal inferences. *Medical engineering & physics*, 31 (6):681–687, 2009.

[17] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134, 2018.

[18] Bo-Lin Jian, Chieh-Li Chen, Wen-Lin Chu, and Min-Wei Huang. The facial expression of schizophrenic patients applied with infrared thermal facial image sequence. *BMC psychiatry*, 17(1):229, 2017.

[19] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018.

[20] Seungryong Kim, Dongbo Min, Bumsub Ham, Seungchul Ryu, Minh N Do, and Kwanghoon Sohn. Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2103–2112, 2015.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[22] Hans Lamecker, Thomas H Wenckebach, and H-C Hege. Atlas-based 3d-shape reconstruction from x-ray images. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pages 371–374. IEEE, 2006.

[23] Taehee Lee and Tobias Hollerer. Multithreaded hybrid feature tracking for markerless augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 15(3): 355–368, 2009.

[24] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *ACM siggraph computer graphics*, volume 21, pages 163–169. ACM, 1987.

[25] Miguel Á Martín-Fernández, Rubén Cárdenes, Emma Muñoz-Moreno, Rodrigo de Luis-García, Marcos Martín-Fernández, and Carlos Alberola-López. Automatic articulated registration of hand radiographs. *Image and Vision Computing*, 27(8):1207–1222, 2009.

[26] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–59, 2018.

[27] T Nakamura, M Sato, and H Kajimoto. semi-automatic scoring method for torticollis by using kinect: 328. *Movement Disorders*, 28:120, 2013.

[28] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 436–445. IEEE, 2018.

[29] Alain Pierret, Y Capowiez, L Belzunces, and CJ Moran. 3d reconstruction and quantification of macropores using x-ray computed tomography and image analysis. *Geoderma*, 106(3-4):247–271, 2002.

[30] Ewa Pietka, Arkadiusz Gertych, Sylwia Pospiech, Fei Cao, HK Huang, and Vicente Gilsanz. Computer-assisted bone age assessment: Image preprocessing and epiphyseal/metaphyseal roi extraction. *IEEE transactions on medical imaging*, 20(8):715–729, 2001.

[31] Thammathip Piumsomboon, Adrian Clark, Mark Billinghurst, and Andy Cockburn. User-defined gestures for augmented reality. In *IFIP Conference on Human-Computer Interaction*, pages 282–299. Springer, 2013.

[32] Jhony K. Pontes, Chen Kong, Anders Eriksson, Clinton Fookes, Sridha Sridharan, and Simon Lucey. Compact model representation for 3d reconstruction. *International Conference on 3D Vision (3DV)*, pages –, 2017.

[33] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*, 2017.

[34] Vijay M Rao and David C Levin. The overuse of diagnostic imaging and the choosing wisely initiative. *Annals of internal medicine*, 157(8):574–576, 2012.

[35] Zaidi Mohd Ripin and Ping Yi Chan. Pathological hand tremor measurementâĂŤchallenges and advances. In *International Conference for Innovation in Biomedical Engineering and Life Sciences*, pages 3–8. Springer, 2017.

[36] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

[37] Ryusuke Sagawa, Hiroshi Kawasaki, Shota Kiyota, and Ryo Furukawa. Dense one-shot 3d reconstruction by detecting continuous regions with parallel line projection. In *2011 International Conference on Computer Vision*, pages 1911–1918. IEEE, 2011.

[38] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. 3-d depth reconstruction from a single still image. *International journal of computer vision*, 76(1):53–69, 2008.

[39] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–98, 2018.

[40] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *European Conference on Computer Vision*, pages 294–310. Springer, 2016.

[41] TurboSquid. Anatomy 3d models, 2019. URL www.turbosquid.com/3d-model/anatomy.

[42] Marie Vidailhet, Emmanuel Roze, and Hyder A Jinnah. A simple way to distinguish essential tremor from tremulous parkinsonâĂŹs disease. *Brain*, 140(7):1820–1822, 2017.

[43] Chengde Wan, Angela Yao, and Luc Van Gool. Hand pose estimation from local surface normals. In *European conference on computer vision*, pages 554–569. Springer, 2016.

[44] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–67, 2018.

[45] Chi Xu and Li Cheng. Efficient hand pose estimation from a single depth image. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3462, 2013.

[46] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. A hand pose tracking benchmark from stereo matching. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 982–986. IEEE, 2017.

[47] Xiong Zhang, Qiang Li, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. *arXiv preprint arXiv:1902.09305*, 2019.

[48] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4903–4911, 2017.