1 Abstract

The application of machine-learning algorithms, notably CNN, for hand pose estimation has enabled significant improvements in accuracy. However, self-occlusion of the fingers, the large variety of poses and viewing angles makes generalisation difficult. This research builds on recent advances in 3D pose estimation to determine the angle of contracture and subsequently classify Dupuytren's disease. A machine-learning algorithm trained on a custom synthetic hand pose dataset is used to estimate a 3D hand model from 2D images of hands affected by the disease. Simple geometry is then used to determine the angle of contracture at each joint and the total angle of contracture. The accuracy of the algorithm is tested by comparing predicted results with actual measurements for the contracture of the hand. A second approach of predicting the angle directly from the images is also tested. The results obtained indicate a moderate level of success in predicting the angle of contracture and pave the way for future research into the topic.

2 Introduction

Dupuytren's disease is a medical disorder of the hand, which causes one or more fingers to bend towards the palm. Its occurrence is associated with the uneven thickening of palm tissue (fascia) which causes lumps (of tissue) to form under the palm skin. These nodules, when sufficiently large, can form cords (of tissue) which may contract and cause the fingers to bend towards the palm (NHS, 2015). Corrective surgery may straighten the affected fingers, but there is a possibility that the fingers may begin to bend again.

This research intends to create a continuous monitoring tool which patients affected by Dupuytren's contracture can use monitor their recovery from and detect any possible reoccurrence of the affliction. The final product will be a software capable of estimating the angle of contracture from two-dimensional (2D) images of the hand.

Estimating the angles at the joints involves modelling a hand from the target image and this dissertation attempts to solve this underlying hand pose estimation problem. In a computer vision context, hand pose estimation is the process of deducing the 3D locations of the hand joint from visual inputs such as RGB images (Barsoum, 2016), This research will focus on single RGB images and not RGB depth images because of the widespread availability of 2D RGB cameras in most modern smartphones.

3 Related Work

That there is no consistent evaluation criterion so comparing the accuracy and performance of various approaches is difficult (Supancic et al., 2015). While Tzionas and Gall (2013) proposed a standardised benchmark for measuring errors using a generalised Chamfer distance it has not been used by most of the reviewed studies.

3.1 Existing Datasets

There are several datasets of hand-poses and their associated joint measurements such as the NYU dataset (Tompson et al., 2014) and the Rendered Hand pose dataset (Zimmermann and Brox, 2017). There is a significant variation (in background and poses) between the various training datasets, and this makes it impossible to objectively compare the various datasets (Supancic et al., 2015). It is interesting to note that there is only a small variation between the several model architectures and according to Supancic et al., (2015) this reinforces the idea that the training data is more important than the approach. Moreover, the datasets tend to make unstated assumptions about the viewpoint and background objects. It has also been observed that the distance of the camera from the hand affects accuracy particularly for depth cameras (Xu and Cheng, 2013).

Synthetic datasets are a possible way to capture a larger variety of poses with more accurate annotations but are unable to consider sensor

features such as noise, and it is complicated to apply anatomical constraints of the hand. Current synthetic datasets can be improved by adding simulated noise and cluttered backgrounds and also modelling hands interacting with each other (Barsoum, 2016). The earlier works focused almost exclusively on real human data and often contained annotation errors, and synthetic datasets are becoming more common (Supancic et al., 2015).

3.2 Hand Pose Estimation

Hand tracking can be divided into two broad categories (De La Gorce et al., 2011):

1. **Discriminative methods** use supervised learning techniques such as regression to reconstruct hand poses from single images. However, since the sample space of the possible positions of the fingers is large, these methods (when used alone) suffer from a significant level of inaccuracy. Nevertheless, the estimated pose can be used to initialise a generative technique and vice-versa.
2. **Generative methods** use a 3D hand model whose pose is aligned with that of the observed image.

Many of the more recent approaches use a combination of both discriminative and generative techniques. This review will consider only discriminative methods as they are they are the focus of this paper.

3.2.1 Discriminative Methods

The use of machine learning techniques to estimate hand poses has been gaining popularity, Convolutional Neural Networks (CNN), in particular, are very popular with researchers. One of the reasons for the surge in popularity is that the networks automatically prioritise features and is less prone to human errors involving feature elimination (Gattupalli et al., 2016). Oberweger et al. (2015) made use of a variety of CNN's to calculate the 3D joint locations from an image depth map. They demonstrated that it is feasible to incorporate partial skeletal information about the hand as prior information for multi-layer CNN. Similarly, Ge et al. (2016) made use of Convolutional Neural Networks (CNN) to generate heat maps with estimates for the locations of the joints. The model could accurately determine hand poses in real time and displayed a significant improvement over a similar work by Tompson et al. (2014). Many of the works described above do not consider the physical limitations of the hand, such as the maximum possible rotation angles, during the network training stage. The techniques tend to post-process the data which can be inefficient and Zhou et al. (2016), building on the approach of Oberweger et al. (2015), fully exploited the skeletal features of the human hand. However, in their study the angle error, though less than previous literature, was significant at 12 degrees.

Convolutional neural networks (CNN), have been used to estimate hand poses either via

1. Regression of heat-maps representing joint likelihoods. These techniques are computationally expensive and need to be constrained by the anatomy of the hand. These techniques use deep learning approaches with multiple iterations and incorporate both generative and discriminative techniques (Chen et al., 2017).
2. Regression of continuous outputs representing joint locations. (Madadi et al., 2017a). For example, Sinha et al. (2016) used a CNN architecture to reduce the dimensionality of depth maps. A pool of activation features was then used with the CNN to estimate the hand pose. These techniques attempt to optimise the estimation by making use of the hand anatomy as prior information (Chen et al., 2017).

Wei et al. (2016) made use of convolutional neural networks to predict poses of the human body from 2D RGB images. Consecutive predictors make dense estimates of a specified body part for each pixel in the picture, enabling the algorithm to locate the different body parts of the human body and thus can be used to model

articulate human body poses. Zimmermann and Brox (2017) leveraged on these pose machines to determine the location of the hand and then used an additional multilayer convolutional network to model the individual joints of the finger. The work by Zimmermann and Brox (2017) uses only single RGB images to determine hand pose and outperforms other techniques for similar non-depth images. Transfer learning was also leveraged by (Gattupalli et al., 2016) who trained the human body pose estimator Deep Pose (Toshev and Szegedy, 2014) on a hand pose dataset and transferred the learned parameters as initial weights for their approach.

Many approaches have used complicated layers *multiple* CNNs to estimate hand poses, and Guo et al. (2017) attempted to reduce this complexity while maintaining accuracy that multiple CNN's provide. The feature maps of the CNN were partitioned into regions before being fed into fully-connected layers, and these areas were used to mimic the behaviour of individual CNN.

CNN's are much slower than Random Decision Forests, and this can impact both performance and cost (Krejov et al., 2017). Moreover, Tree based approaches have risen in popularity because these approaches are not algorithmically complex and have a quick recognition time (Kirac et al., 2014).

In an attempt to create a standardised technique of evaluating the various approaches Supancic et al. (2015) ran multiple algorithms across various datasets and observed that deep methods tend to generalise well for uniform backgrounds. They also noted that the accuracy increases with training sample size, but at an exponential performance cost, suggesting that current hardware may be unable to handle a large synthetic dataset.

Major disadvantages to using discriminative approaches are that the hand has a significant number of viewing points and the hand can self-occlude in numerous ways. An exponential amount of training data is required to cater for these drawbacks (Tang et al., 2017). Banzi et al. (2016) attempted to account for this self-occlusion by using a probabilistic model to predict the position of occluded fingers. There have also been attempts by Madadi et al. (2017) to solve the self-occlusion problem by matching the shape of the hand pose with possible candidates. The researchers divided the pose estimation problem into two sub problems – palm pose estimation and hand pose estimation. However, they used a custom data set so the performance of their approach cannot be reasonably compared with other methods. Chen et al. (2017) attempted to minimise the effect of self-occlusion by using a spherical coordinate system that uses joints with the smallest error to define the axes. This approach located the hand using the CNN in Chen et al. (2016) and generated a custom dataset by rotating images of existing datasets.

4 Methodology

A set of CNN's based on the ColorHandPose3D network (Zimmermann and Brox, 2017) will be trained on the dataset mentioned above to be able to predict better hand poses of individuals affected by Dupuytren's disease. The network will be initialised using the Zimmermann and Brox (2017) weights. The algorithm will first estimate the location of the hand using a Convolutional Pose Machine (Wei et al., 2016). The algorithm will then crop and magnify the location of the hand after which additional networks will generate score maps for the likely locations of the joints on the cropped image as well as estimated 3D coordinates of the joints. These score maps will be used to model the skeleton of the hand from the captured images.

The synthetic dataset in will be divided into a training set and a testing set consisting of 1855 and 500 images respectively.

4.1 Convolutional Neural Network

4.1.1 Training

The convolutional neural network will incorporate concepts of transfer learning and will be initialised using the weights of the

ColorHandPose3DNetwork. A Tensorflow session will then run with the training images, and the ground truth-values will be fed into the network.

The training process involves the following steps:

1. Load initial weights
2. Read training images
3. Process training images
4. Read training ground truth values
5. For each CNN in Tensorflow run the training image
6. Save the weights after training

4.1.1.1 Test Algorithm

The parameters of the trained classifier will be embedded in software, which will process the target image in the following order:

1. **Pre-process image.** This will involve scaling the image to a size usable by the pose machine.
2. **Use the convolutional pose machine to locate the hand.** The image will then be cropped, and the cropped image will be used in the following steps
3. **Use a CNN to determine if the hand is a left or right hand**. If the hand is a right hand, the modeller will mirror the hand for estimation
4. Use a CNN to predict the location of the joints. This will be used to model the hand
5. **Calculate the angles of interest from the model** using simple geometry as per the literature review.

The algorithm will utilise on three CNN's which each perform different tasks:

4.1.1.2 HandSegNet

This network will locate the hand using skin-colour and crop the source image so that the entirety of the detected hand is in the image. The network is the same as that of Zimmermann and Brox (2017) and consists of 16 convolutional layers with a ReLU activation function.

4.1.1.3 PoseNet

PoseNet predicts score maps for the 2D locations of the coordinates, where each score map represents the likelihood of occurrence of each joint. The network uses an encoder-decoder architecture which is based on the one proposed by Wei et al. (2016), and the network was initialized using weights by s Zimmermann and Brox (2017) and further trained on the synthetic training dataset. The network consists of three sets of convolution layers.

4.1.1.4 PosePrior

Given a score map from PoseNet, the PosePrior network predicts the possible 3D coordinates. It was trained with prior information on the hand anatomy. Simply put it will predict the likely 3D coordinates of a joint given the 2D coordinates (from the score map).

4.2 Second Approach of Predicting Angle Directly

Instead of modelling a hand from RGB images, a secondary approach of predicting the angle directly from images of hands affected by Dupuytren's contracture was adopted. This approach made use of MATLAB and the Neural Network Fitting tool and consisted of the following training steps:

1. Resizing the image to 256x256 pixels
2. Converting the colour image to a grayscale one
3. Merging the rows in the image (to simplify training)
4. Training a scaled conjugate backpropagation neural network with 10 hidden layers on the pre-processed image and the actual angle of contracture as the target. The network was trained using a mean square error function for 93 Epochs

The images were tested by pre-processing them, running the network on them and comparing the estimated angle with the actual angle. The architecture of the neural network is shown in the diagram below.

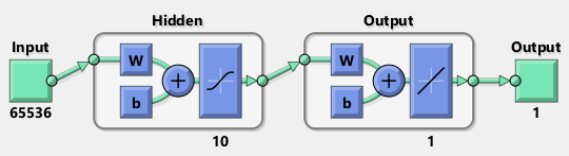

Figure 1: Architecture of Network Used in Second Approach

5 Results & Findings

The algorithm was initially run on 100 test images of the synthetic dataset, the predicted angles were compared to the actual angles, and the results are shown below

The RMSE in estimating the total angle of contracture is 22°. The hand modelling approach was observed to have a bias towards overestimating the angle of contracture by around 13.65°. This bias is evident on the box and whisker chart shown below.

After creating a bias adjustment factor for the overestimation, the updated algorithm was run against a larger test dataset consisting of 500 images.

Table 1: Results for Test Case 1

Finger	Finger Joint	Root Mean Square Error (Degrees)
Index	DIP	13.07112175
	PIP	17.72014119
	MCP	14.62630448
Middle	DIP	14.77060114
	PIP	15.34358672
	MCP	27.99771122
Ring	DIP	17.97389988
	PIP	14.63741393
	MCP	20.65981877
Pinky	DIP	22.0765673
	PIP	15.7370576
	MCP	14.75136762
Total Angle of Contracture		22.0368

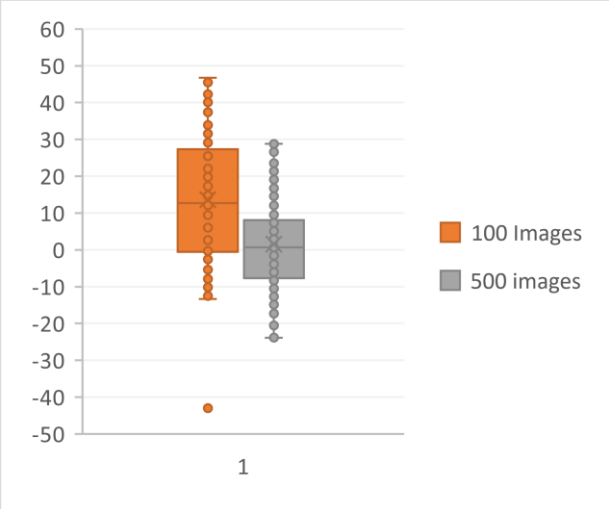

Figure 2: Box & Whisker Diagram of Error for test before (100 images) and after bias adjustment (500 images)

The results from running against the 500 test images indicated a much smaller overestimation of 1.6° compared to the 15.2° for the unadjusted network. The box and whisker diagram below indicates that while there was still an overestimation it the values were closer to the true values. It was noted that the standard deviation of the error was 12°, which is significant as this means the algorithm on average would underestimate or overestimate, by that value.

6 Discussion of Results

The algorithm when tested on the synthetic dataset overestimated the angle of contracture. This is attributable to the fact that the 3D hand modeller inclined to model the fingers as more bent than they were. In fact, during the tests the angle of contracture was overestimated in 74% of the cases. This may be because the training images did not include sufficient negative samples i.e. images of the hand with no contracture. Nevertheless, after introducing a bias adjustment factor the accuracy of the algorithm improved on a larger test set. The results obtained were less accurate than expected and this can be attributed to the small size of the training dataset.

The second approach that predicted the angles directly had a much lower RMSE of less than two. However, given the small size of the training dataset it is more than likely that the neural network over-fit to the synthetic dataset. A much

larger dataset would have been more suitable for training the network as suggested by Supancic et al. (2015).

The accuracy of both approaches can be improved by increasing the size and variety (in pose, background) of the training dataset and this idea is also supported by the work done by Barsoum (2016). Incidentally, the algorithm performed better on hands with a lighter skin colour and the skin colour on the synthetic dataset can be varied to achieve less discriminatory results. The evaluation datasets may not represent all the hand poses possible from people affected by Dupuytren's Finger and a larger dataset with a larger pose variety can be used to more realistically estimate the accuracy of the algorithm.

The study used artificial intelligence concepts because of the comparative success such techniques have obtained in the computer vision problems. While, the results were far from perfect there was a general indication that the algorithm was somewhat accurate with the accuracy being sufficient to classify the disease under the Tubiana classification system.

7 Conclusion

This software has potential use in aiding doctors in the measurement of the angles of contracture. The results of the study itself would also lay a sufficient baseline from which other researchers can continue to investigate Dupuytren's contracture. The research enables continuous monitoring for patients who have received treatment allowing them to detect any resurgence of the disease. Different physicians often obtain different measurements of the angle of contracture and this research can be used to assist in establishing a consensus on the actual angle.

The results obtained from the dissertation enables further research into the topic, which would see more cost effective and accurate assessment of Dupuytren's finger. The focus of future work would be on improving the accuracy of the hand pose estimation (and hence the angle of contracture) by training on a larger training dataset with a much wider variety of poses and background. Ideally, real pictures of actual hands affected by Dupuytren's contracture could be used to measure the angle of contracture. Eventually, depth cameras will become a standard feature of smartphones and this research can be adapted to leverage on depth images, which would significantly improve the accuracy.

8 References

Banzi, J.F., Zhongfu Ye, Bulugu, I., 2016. A novel hand pose estimation using dicriminative deep model and Transductive learning approach for occlusion handling and reduced descrepancy. In: 2016 2nd IEEE International Conference on Computer and Communications (ICCC). IEEE, pp. 347–352.

Barsoum, E., 2016. Articulated Hand Pose Estimation Review 1–50.

Chen, T.-Y., Wu, M.-Y., Hsieh, Y.-H., Fu, L.-C., 2016. Deep learning for integrated hand detection and pose estimation. In: 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, pp. 615–620.

Chen, T., Ting, P.-W., Wu, M., Fu, L., 2017. Learning a deep network with spherical part model for 3D hand pose estimation. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 2600–2605.

De La Gorce, M., Fleet, D.J., Paragios, N., 2011. Model-based 3D hand pose estimation from monocular video. IEEE Trans. Pattern Anal. Mach. Intell. 33, 1793–1805.

Gattupalli, S., Ghaderi, A., Athitsos, V., 2016. Evaluation of Deep Learning based Pose Estimation for Sign Language Recognition. arXiv cs.CV 2, 9065.

Ge, L., Liang, H., Yuan, J., Thalmann, D., 2016. Robust 3D Hand Pose Estimation in Single Depth Images : from Single-View CNN to Multi-View CNNs. In: Cvpr.

Guo, H., Wang, G., Chen, X., Zhang, C., 2017. Towards Good Practices for Deep 3D Hand Pose Estimation.

Kirac, F., Kara, Y.E., Akarun, L., 2014. Hierarchically constrained 3D hand pose estimation using regression forests from single frame depth data. Pattern Recognit. Lett. 50, 91–100.

Krejov, P., Gilbert, A., Bowden, R., 2017. Guided optimisation through classification and regression for hand pose estimation. Comput. Vis. Image Underst. 155, 124–138.

Madadi, M., Escalera, S., Baro, X., Gonzalez, J., 2017a. End-to-end Global to Local CNN Learning for Hand Pose Recovery in Depth data.

Madadi, M., Escalera, S., Carruesco, A., Andujar, C., Baro, X., Gonzalez, J., 2017b. Occlusion Aware Hand Pose Recovery from Sequences of Depth Images. 2017 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG 2017) 230–237.

NHS, 2015. Dupuytren's contracture - NHS Choices [WWW Document]. URL http://www.nhs.uk/conditions/Dupuytrens-contracture/Pages/Introduction.aspx (accessed 6.7.17).

Oberweger, M., Wohlhart, P., Lepetit, V., 2015. Hands Deep in Deep Learning for Hand Pose Estimation.

Sinha, A., Choi, C., Ramani, K., 2016. DeepHand: Robust Hand Pose Estimation by Completing a Matrix Imputed with Deep Features. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4150–4158.

Supancic, J.S., Rogez, G., Yang, Y., Shotton, J., Ramanan, D., 2015. Depth-Based Hand Pose Estimation: Data, Methods, and Challenges. 2015 IEEE Int. Conf. Comput. Vis. 1868–1876.

Tang, D., Chang, H.J., Tejani, A., Kim, T.-K., 2017. Latent Regression Forest: Structured Estimation of 3D Hand Poses. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1374–1387.

Tompson, J., Stein, M., Lecun, Y., Perlin, K., 2014. Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks. ACM Trans. Graph. 33, 1–10.

Toshev, A., Szegedy, C., 2014. DeepPose: Human Pose Estimation via Deep Neural Networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1653–1660.

Tzionas, D., Gall, J., 2013. A Comparison of Directional Distances for Hand Pose Estimation. pp. 131–141.

Wei, S.-E., Ramakrishna, V., Kanade, T., Sheikh, Y., 2016. Convolutional Pose Machines. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 4724–4732.

Xu, C., Cheng, L., 2013. Efficient Hand Pose Estimation from a Single Depth Image. In: 2013 IEEE International Conference on Computer Vision. IEEE, pp. 3456–3462.

Zhou, X., Wan, Q., Zhang, W., Xue, X., Wei, Y., 2016. Model-Based Deep Hand Pose Estimation. Proc. 25th Int. Jt. Conf. Artif. Intell. (IJCAI 2016) 2421–2427.

Zimmermann, C., Brox, T., 2017. Learning to Estimate 3D Hand Pose from Single RGB Images. Cvpr.