Extending Convolutional Pose Machines for Facial Landmark Localization in 3D Point Clouds

Eimear O' Sullivan
Imperial College London
e.o-sullivan16@imperial@ac.uk

Stefanos Zafeiriou
Imperial College London
s.zafeiriou@imperial.ac.uk

Abstract

In this work we address the problem of landmark localization in 3D point clouds by extending the convolutional pose machine (CPM) architecture to facilitate landmark localization in 3D point clouds. Making use of PointNet++, we are able to construct an architecture that is invariant to the ordering of an input point cloud. The sequential CPM architecture facilitates allows initial heatmaps to be iteratively refined in a series of point convolutional stages to yield robust landmark predictions. We propose to evaluate our approach for 3D facial landmark localization on benchmark face databases, BU-3DFE, BP4D-Spontaneous and BP4D+. The robustness of the approach to the size of the input point cloud will be assessed, and the contribution of the CPM stages will be evaluated in an ablation study.

1. Introduction

With the recent advancements in 3D capture technologies, the availability of 3D data in the form of meshes and point clouds has become ever more prevalent. With this, 3D landmark localization has become an increasingly studied topic, and has been applied in a diverse range of fields including face verification, facial expression recognition, facial alignment and morphometric analysis.

Many 2D landmark localization approaches have benefited from the use of heatmaps to accurately encode the likelihood of a landmark occurring at a given location [6, 11, 18]. The use of heatmaps has also been successfully applied to the prediction of 3D landmarks from 2D images, both for the face [2, 19] and body [8, 10]. Many of these make use of residual or stacked hourglass networks to refine the predicted heatmaps and improve landmark localization accuracy. One such architecture is the convolutional pose machine (CPM), a sequential heatmap prediction framework that enables increasingly refined landmark predictions and has been used to achieve state-of-the-art results in face and body landmark localization [6, 18].

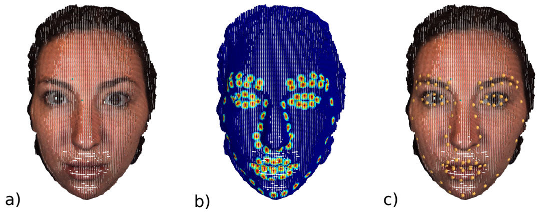

Figure 1. Overview: a) Initial point cloud, b) Predict heatmaps, c) Landmark localization. Colour shown for visualization only.

Given the success of the approach in these domains, we hypothesize that the extension of heatmaps, and CPMs in particular, could lead to substantial gains for the prediction of 3D landmarks from 3D point clouds. As methods for processing unordered point clouds have improved substantially in recent years [5, 13, 20, 22], this concept has become increasingly feasible. We aim to make the following contributions: a) extend the CPM architecture for landmark localization in 3D point clouds, and b) quantitatively evaluate the proposed approach via an ablation study and a comparison to current state-of-the-art in 3D landmark localization.

2. Related Work

Many 3D approaches to landmark localization have exploited the strength of 2D techniques by rendering images of a textured mesh from multiple viewpoints and projecting detected keypoints onto the 3D space of the mesh [1], however these approaches can be sensitive to illumination, pose and expression. Others have combined texture-based information with spatial information by fitting an active appearance model (AAM) to intensity and depth maps of a surface [7]. In [4] ensemble landmarking is used to coalesce extracted features from texture, depth and height maps.

Approaches that consider only the geometric structure have also been proposed. Wang *et.al.* [17], convert 3D data to attribute maps such as intrinsic curvature, normals and depth and use these to train a fused Convolutional Neural Network (CNN). Sun *et.al.* made use of vertex-flow to create an AAM for landmark tracking [16]. In [15], an architecture for 3D facial annotation based mesh shape is proposed. Curvature analysis is used to detect fiducial points, which are used to initialize the remaining landmarks via an Active Normal Map (ANM), prior to a final iterative refinement stage. This approach relies on handcrafted features however, namely the concavity and the convexity of the eye corners and nose tip respectively, it is not readily transferable to other domains. A model-based approach is also used in [3], where curvature about the landmarks is used to create a shape index-based statistical shape model (SI-SSM).

In the context of point cloud processing, a number of deep learning approaches that directly consume point clouds have been proposed. The PointNet++ architecture has demonstrated great success for classification and segmentation of point clouds [14]. They propose point convolutions, which are invariant to the ordering of input points, and makes use of a sampling and grouping strategy for pooling. PointNet [13] also been successfully applied to find correspondences between sets of point clouds [5]. Other approaches for 3D feature detection in point clouds [20] and voxels have also been proposed [22] for the purpose of 3D scene alignment.

3. Methodology

A point cloud, $P = [x_1^T, x_2^T, ..., x_n^T]$, is defined as a set of n 3D points, $x_i = [x_{ix}, x_{iy}, x_{iz}]^T$. The set of m, landmarks, L, is similarly defined.

The heatmaps, h, are constructed by applying a Gaussian peak at the ground truth landmark location. A 1D heat vector, $h = [h_1, h_2, ..., h_n]^T$, is constructed for each of the m landmarks, where $h_i \in \mathbb{R}$ is the value associated with vertex x_i in the mesh based on its proximity to the landmark. The set of all heatmaps is denoted as H.

3.1. Landmark Localization

The proposed architecture follows that of previous work on CPMs [6, 18], which iteratively refine landmark predictions in a series of successive convolutional stages. This architecture is outlined in Figure 2. The feature extraction block is implemented using PointNet++ [14]. The initial set abstraction layers of the architecture perform grouping and pooling, while feature propagation is facilitated by skip link concatenations. For full details, refer to [14]. The output layer of this block is modified to gives H_0, the initial heatmap estimates.

Three subsequent stages are used for refinement. Stage 1 takes H_0 as input and outputs heatmap H_1. Stage 2, s_2, and stage 3, s_3, take the output of the previous stage concatenated with the output of the feature extraction block as input, as shown in Eq. 1. To calculate the final landmark predictions, the maximum three points in each h_i are chosen. The i^{th} landmark is then calculated as the barycentre of the three corresponding vertices.

$$\begin{aligned} s_2(H_0, H_1) &= H_2 \\ s_3(H_0, H_2) &= H_3 \end{aligned} \tag{1}$$

The loss function minimized during training is the sum of the mean squared error for the output heatmaps at each of the three prediction stages, z

$$Loss = \sum_{i=1}^{m} \sum_{z \in 1,2,3} \|h_i(z) - h_i^*(z)\|^2 \tag{2}$$

where h is the ideal heat map, and h^* is the corresponding prediction. The model is constructed with Pytorch [12], uses a batch size of 8, and Adam optimization [9] with an initial learning rate of 0.001. All meshes are normalized in a pre-processing step. During training, the point clouds, and their corresponding landmarks, are randomly rotated about the x, y and z axes as a means of data augmentation.

4. Experimental Protocol

We propose to evaluate our approach on the BU-3DFE [21], the BP4D-Spontaneous (BP4D-S) [23] and BP4d+ [24] databases. BU-3DFE contains the scans of 100 individuals, while BP4D-S consists of 41 subjects. Both have previously been used to provide a benchmark for 3D landmark localization. BP4D+ contains the scans of 140 individuals. All databases include both male and female subjects from a wide range of ethnicities, displaying a variety of emotions, and are annotated with 83 facial landmarks.

Precision and absolute error will be used to evaluate the proposed approach. Precision rate is the proportion of predicted landmarks within a specified range from the ground truth landmarks, while absolute error refers to the Euclidean distance between predicted landmark and the ground truth.

4.1. Ablation Study

The proposed architecture consists of three fully convolutional stages in which the landmark heatmaps are predicted and further refined. In this section we aim to quantify the contribution of each of these stages via an ablation study by evaluating the accuracy of the landmarks produced by the heatmaps at each of the three stages.

4.2. Comparison with State-of-the-Art

The precision rate will be evaluated using the BU-3DFE database, following the precedence in [7, 15]. 80 of the 100 individuals in the database will be randomly selected to form the training set, while the remaining 20 individuals will comprise the test set. The experiment will be repeated 5 times, with a different test set each time, so that each individual is assigned to the test set on one occasion. The

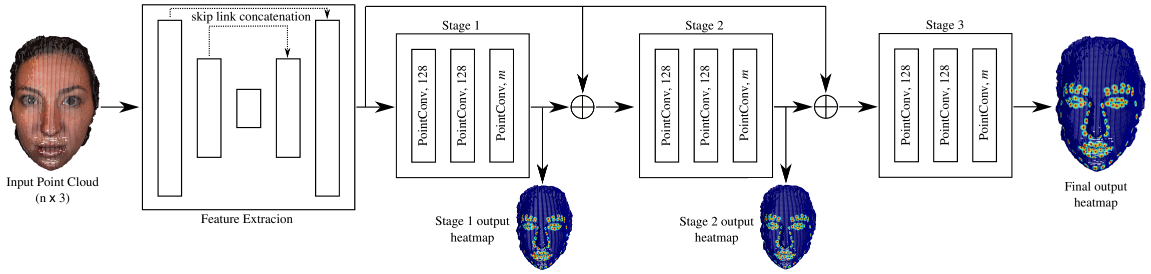

Figure 2. Network architecture. PointNet++ architecture is used for initial feature extraction, followed by three point convolutional layers. The numbers indicating the quantity of output layers in the convolutional stages. ⊕ signifies a concatenation operation.

mean precision and absolute error over all experiments will be reported and compared to those of [3, 7, 15]

Accuracy on BP4D-S will be compared to [3, 16] using the mean square error (MSE) between ground truth and predicted landmarks. The one-point spacing procedure outlined in the same paper will be used, where the one-point spacing is defined as the distance between the closest pair of points in the 3D scan ($\approx 0.5mm$).

4.3. Effect of Point Cloud Size

Finally, the performance of the model for a different input sizes will be evaluated, for both landmark localization accuracy and processing speed of a single point cloud.

References

[1] J. Booth et al. A 3d morphable model learnt from 10,000 faces. In *CVPR 2016*, pages 5543–5552, 2016. 1

[2] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). pages 1021–1030. IEEE, Oct 2017. 1

[3] S. Canavan, P. Liu, X. Zhang, and L. Yin. Landmark localization on 3d/4d range data using a shape index-based statistical shape model with global and local constraints. *CVIU*, 139:136–148, Oct 2015. 2, 3, 4

[4] M. A. de Jong et al. Ensemble landmarking of 3d facial surface scans. *Scientific reports*, 8(1):12, Jan 8, 2018. 1

[5] H. Deng, T. Birdal, and S. Ilic. PPF-FoldNet: Unsupervised learning of rotation invariant 3d local descriptors. Aug 30, 2018. 1, 2

[6] X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Style aggregated network for facial landmark detection. pages 379–388. IEEE, Jun 2018. 1, 2

[7] G. Fanelli, M. Dantone, and L. Van Gool. Real time 3d face alignment with random forests-based active appearance models. pages 1–8. IEEE, Apr 2013. 1, 2, 3, 4

[8] S. Kinauer, R. A. Guler, S. Chandra, and I. Kokkinos. Structured output prediction and learning for deep monocular 3d human pose estimation, Nov 1, 2017. 1

[9] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. Dec 22, 2014. 2

[10] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. Mar 22, 2016. 1

[11] G. Papandreou et al. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. Mar 22, 2018. 1

[12] A. Paszke et al. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017. 2

[13] C. R. Qi, H. Su, M. Kaichun, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. pages 77–85. IEEE, Jul 2017. 1, 2

[14] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Jun 7, 2017. 2

[15] J. Sun, D. Huang, Y. Wang, and L. Chen. Expression robust 3d facial landmarking via progressive coarse-to-fine tuning. *TOMM*, 15(1):1–23, Feb 25, 2019. 2, 3, 4

[16] Y. Sun, X. Chen, M. Rosato, and L. Yin. Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 40(3):461–474, May 2010. 2, 3, 4

[17] K. Wang, X. Zhao, W. Gao, and J. Zou. A coarse-to-fine approach for 3d facial landmarking by using deep feature fusion. *Symmetry*, 10(8):308, Aug 1, 2018. 1

[18] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. pages 4724–4732. IEEE, Jun 2016. 1, 2

[19] P. Xiong, G. Li, and Y. Sun. Combining local and global features for 3d face tracking. pages 2529–2536. IEEE, Oct 2017. 1

[20] Z. J. Yew and G. H. Lee. 3dfeat-net: Weakly supervised local 3dfeatures for point cloud registration. Jul 24, 2018. 1, 2

[21] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. pages 211–216. IEEE, 2006. 2

[22] A. Zeng et al. 3DMatch: Learning local geometric descriptors from RGB-D reconstructions. pages 199–208. IEEE, Jul 2017. 1, 2

[23] X. Zhang and others. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, Oct 2014. 2

[24] Z. Zhang et al. Multimodal spontaneous emotion corpus for human behavior analysis. pages 3438–3446. IEEE, Jun 2016. 2

	Abs. Err. (mm)		Mean Precision (%)	
Layer	mean	SD	$< 5mm$	$< 10mm$
1	5.18	4.58	60.09	91.83
2	4.78	4.19	63.96	93.89
3	4.73	4.26	63.77	93.78

Table 1. Effect of CPM layers on landmark prediction accuracy.

	Abs. Err. (mm)		Mean Precision (%)	
	mean	SD	$< 5mm$	$< 10mm$
Ours	3.97	2.39	73.30	97.17
Fanelli [7]	4.22	2.99	72.51	95.80
Sun [15]	3.47	2.95	84.13	96.96

Table 2. Comparison on the BU-3DFE Database for the 14 keypoints outlined in [7].

	Ours	Cavanan [3]	Sun [16]
MSE	5.64	9.6	2.54

Table 3. MSE for BU-3DFE. A unit error of 0.5 mm is used.

		BU-3DFE		BP4D+	
Pts	fps	mean	SD	mean	SD
4096	5.18	4.75	3.37	5.39	4.82
2048	13.07	4.99	3.09	4.24	2.57
1024	23.04	5.15	2.95	4.89	3.98
512	27.81	5.85	3.24	5.68	3.38

Table 4. Effect of point density on absolute error (mm) and processing speed in frames per second (fps).

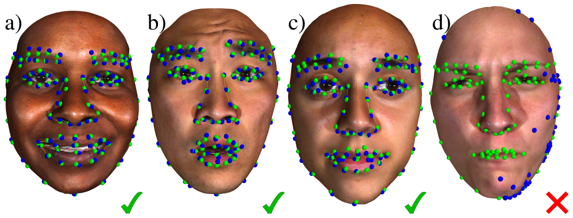

Figure 3. Sample landmark predictions at 4096 points. Texture and surface shown for visualisation purposes only. Predicted landmarks are in blue, while ground truth are in green.

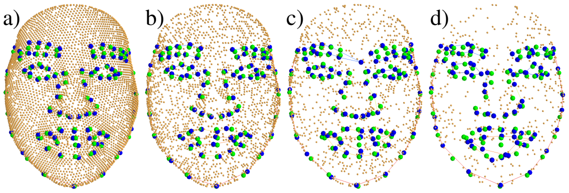

Figure 4. Predicted landmarks for the same point cloud at different point densities; a) 4096, b) 2048, c) 1024, and d) 512.

5. Results

5.1. Ablation Study

The results of the ablation study, which allow for the contribution of each convolutional stage to be quantified, are outlined in Table 5.1. Values were calculated for the BU-3DFE dataset at a point density of 4096. The results indicate that the refinement stages do allow for some improvement on the accuracy of landmark prediction. Beyond the second stage, however, this improvement is minimal.

5.2. Comparison with State-of-the-Art

In lieu of evaluation on BP4D-S, all comparisons were made with BU-3DFE. Again, 4096 points were used for evaluation. For approximately 3% of cases, the network failed to predict coherent landmarks, as shown in the example in Figure 3. These cases were omitted from the analysis.

Table 5.2 compares our results with those of the 14 fiducial landmarks specified in [7, 15]. The mean absolute error and precision rate for each of these points in all test samples are reported. Table 5.2 reports the unit MSE for the 83 landmark points as compared with the results of [3] and [16]. While our network does not top the best performing system in either case, acceptable results are nonetheless achieved.

5.3. Effect of Point Cloud Size

Table 5.3 summarises the effect of point cloud density on the accuracy of landmark predictions and the processing speed for BU-3DFE and BP4D+. Predictably, as the point cloud density decreased, so too did the landmark localisation accuracy. The exception to this is BP4D+ at 4096, where the network appeared to fit poorly to the training data. The failure rate remained consistent at all point densities.

The processing time for an input point cloud decreased with the point density. At 4096 points per sample, the average processing time was observed to be 193.11 ms, giving a frame rate of 5.18 fps. Compared with [3, 7, 15], which achieve a approximately 3, 25 and 0.71 fps respectively, our system network performs respectably in this area.

6. Conclusion

In this paper we have presented a method for landmark localization in point clouds. The use of fully convolutional refinement stages allowed for increased landmark localisation accuracy. The greater the number of points in the input data, the more accurate the landmark predictions were found to be. Although this approach does not improve upon the state-of-the-art, respectable performance is achieved. Future work will aim to improve the landmark localisation accuracy and reduce the failure rate by exploiting the statistical positional relationships between landmarks.