

An automated CNN-based 3D anatomical landmark detection method to facilitate surface-based 3D facial shape analysis

Ruobing Huang[★], Michael Suttie[◇], J.Alison Noble[★]

[★]Institute of Biomedical Engineering, University of Oxford

[◇] Nuffield Department of Women’s and Reproductive Health, University of Oxford

Abstract. Maternal alcohol consumption during pregnancy can lead to a wide range of physical and neurodevelopmental problems, collectively known as fetal alcohol spectrum disorders (FASD). In many cases, diagnosis is heavily reliant on the recognition of a set of characteristic facial features, which can be subtle and difficult to objectively identify. To provide an automated and objective way to quantify these features, this paper proposes to take advantage of high-resolution 3D facial scans collected from a high-risk population. We present a method to automatically localize anatomical landmarks on each face, and align them to a standard space. Subsequent surface-based morphology analysis or anatomical measurements demands that such a method is both accurate and robust.

The CNN-based model uses a novel differentiable spatial to numerical transform (DSNT) layer that could transform spatial activation to numerical values directly, which enables end-to-end training. Experiments reveal that the inserted layer helps to boost the performance and achieves sub-pixel level accuracy.

Keywords: FASD, landmark detection, 3D facial analysis

1 Introduction

Maternal drinking in pregnancy can result in fetal central nervous system damage, and fetal and young child growth deficiency, cranio-facial abnormalities, learning disabilities and functional impairments [1]. The range of phenotypes varies greatly in severity and presentation, largely dependent on the level, pattern, and timing of maternal alcohol consumption therefore termed as: fetal alcohol spectrum disorder (FASD). The prevalence of FASD has been documented as high as 13.5% - 20.8% in some regions, and produces an immense burden to families and society [1].

Formal diagnosis at the earliest possible stage is paramount, as it allows early intervention and reduces the risk of secondary disabilities [1]. However, the diagnosis of FASD is particularly challenging due to the inconspicuous brain malformations in young children, unreliable reports of maternal alcohol usage, and

misdiagnosis of syndromes with similar characteristics. Recognition of the most severe form, fetal alcohol syndrome (FAS), is reliant on the identification of 3 cardinal facial features: short palpebral fissure length (PFL), a smooth philtrum and a thin upper lip. Current methods for facial assessment are necessarily subjective and vary between clinicians. For example, clinicians score the lip and the philtrum using a 5-point Likert scale for visual comparison (Fig. 1) [1]. To measure PFL, some clinicians place a ruler on the subjects’ face, whereas others use a semi-automated 2D software. Both methods for facial evaluation involve subjectivity, and accuracy can be influenced by the skill and experience of the clinician. Some prior studies have pointed out that PFL obtained in 3D can be more stable while they relied on manually annotated 3D landmarks [2].

To overcome these difficulties, we propose an automated method to detect 3D anatomical landmarks to calculate anthropometric measurements and facilitate surface-based 3D shape analysis. Using a differential spatial-to-numerical layer, the trained model is compact and predicts the target coordinates directly without the need for post-processing or fully-connected layers. To the best of our knowledge, this is the first time that a fully automatic method has been proposed to assist FASD diagnosis using 3D facial form.



Fig. 1: FAS Diagnosis ¹



Fig. 2: 3D mesh and the corresponding texture map (zoomed in the eye region). Key part of face is masked to protect privacy.

2 Related works

Before discussing the related literature and the proposed method, we first introduce our dataset as it is different from commonly used medical imaging data. The dataset was collected using a static stereo-photogrammetric camera system produced by 3DMD. By simultaneously taking photos using two cameras with fixed positions, the system is able to reconstruct a dense 3D surface. Each facial scan consists of a 3D mesh (Fig. 2) and two 2D photographs (also referred to as texture maps). The 3D mesh is defined by an OBJ file where the link between each polygon face and the corresponding texture map regions is also included for color rendering. Each 3D mesh consists of more than 30,000 vertices representing surface geometry, and all texture maps are resized to 256×256 pixels. For this

¹ Image credit: <https://www.youtube.com/watch?v=044Zxy3ou8>.

study, 3D images of an ethnically admixed population of unexposed and prenatal alcohol-exposed infants at 1-12 months were used from the Prenatal Alcohol and SIDS and Stillbirth (PASS) Network in Stellenbosch, South Africa (n=777, Fig.3). Among them, 305 are alcohol-exposed cases. The dataset was randomly separated into training (n=622) and test (n=155) sets. 20 reliable landmarks were manually annotated in 3D on each face by an expert, and were projected to each corresponding 2D texture map as training labels for the CNN models.

Manual 3D landmark placement is a tedious and potentially error-prone task, but is often a necessity for 3D shape analysis. Literature is abundant in detecting landmarks in 2D images and relatively new in 3D. We refer the interested reader to [3] for a recent review. Detecting 3D landmarks automatically is difficult as processing a high-resolution 3D mesh is non-trivial. There are also some successful works in computer vision that process 3D points cloud or meshes directly [4–6]. However, these approaches mainly addressed classification or segmentation. More importantly, they usually require a large dataset to train those complicate models which is not available in our case. Another key observation is that the surface of the mesh is less well-defined than the corresponding texture map in regions of complex geometry such as the eye corners and eyebrows (Fig. 2). Therefore, we approach the problem by detecting landmarks on the 2D texture map and propagate the results back to 3D for further analysis.

Besides the common challenges faced in detecting 2D landmarks, e.g. large appearance variations, environmental conditions changes, and occlusions caused by extreme head poses, our task has additional challenges. The age group and ethnic background of the subjects is substantially different from the web-collected datasets (e.g. CelebA) widely used in computer vision. Our preliminary experiments showed that models trained on these datasets failed to detect the face in our dataset constantly (up to 24%)². Further, the annotation of the landmarks is performed in 3D, thus the training labels of our 2D detection model are obtained by projecting these 3D points on the 2D texture map using UV unwrapping. As a result, the location of the same landmark (e.g. the outer corner of the left eye) could appear at very different locations on the texture map and the number of annotated landmarks on each photo is uncertain. The addressed task requires quantitative accuracy, and therefore needs specialized design.



Fig. 3: Examples of subjects’ texture maps contained in this challenging dataset.

² This is echoed by an extensive study in face recognition [7], which showed CV models could fail on nearly 35% on darker skinned females.

Among existing landmark detection methods [8, 9], some of the most successful ones leveraged the power of convolution neural networks (CNNs), and could be mainly divided into two groups: 1) coordinate regression by attaching a fully connected layers at the end; and 2) matching synthetic heatmaps that are generated by placing a 2D Gaussian at the target location. The first approach can be trained using the raw labels while some valuable spatial information is lost during flattening. The heatmap matching approach takes advantage of the translation invariant characteristics of CNNs as neurons can be activated anywhere in the visual field and scored the highest accuracy in many benchmarks. However, during inference, the coordinates are derived by extracting the brightest pixel in the output, which typically involves a non-differential argmax operation. As a result, there is a disconnect between the training loss function and the evaluation metrics that might lead to convergence at a sub-optimal minimum.

3 Method

Figure. 4 shows a schematic of the analysis pipeline. During test time, two unseen texture maps are individually passed into a CNN model for landmark detection. The results are then combined and projected back to automatically label the corresponding 3D mesh. Using the detected set of landmarks we can calculate relevant anatomical measurements, and utilize tools for surface-based shape analysis. By doing so, individual and group dysmorphism can be quantified by calculating the normalized surface-based differences against controls [10].

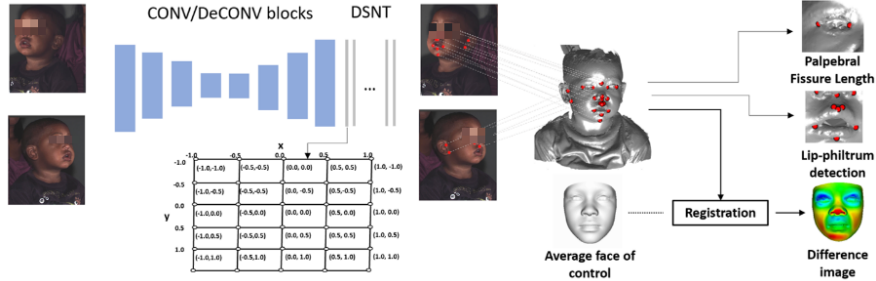


Fig. 4: Schematic of the proposed pipeline. The texture maps are passed into the CNN model which consists of several convolutional (CONV) and deCONV blocks (blue) and DSNT layers (grey) that yield landmark locations (red dots) directly. The results are propagated to 3D to produce clinical measurements and morphological analysis.

DSNT Layer The key idea of the proposed CNN model is the incorporation of a novel layer. As discussed before, the heatmap matching method has one disadvantage that could hamper performance: a gap between the target metric and the optimized loss function. The solution for this boils down to find a differentiable way to extract coordinates from heatmap-like activations.

Following [11], we insert a DSNT layer at the end of our CNN model. The key insight here is that the value of each pixel $p_{i,j} \in \mathcal{P}$ in the predicted heatmap \mathcal{P} essentially indicates the likelihood of that pixel being the target landmark $C = \{x, y\}$. In other words, if we first normalize \mathcal{P} to $\hat{\mathcal{P}} = \frac{\mathcal{P}}{\sum \mathcal{P}}$ such that for $\hat{p}_{i,j} \in \hat{\mathcal{P}}$ it satisfies that $\sum_{i,j} \hat{p}_{i,j} = 1$, $\hat{\mathcal{P}}$ could be interpreted as the probability distribution of the location of the target landmark. The expected value of the target coordinate distribution can then be calculated as the inner product of the normalized heatmap $\hat{\mathcal{P}}$ and the same size coordinate mesh-grid. More formally, given the size of final activation map $\hat{\mathcal{P}}$ is $H \times W$, we construct a 2D mesh-grid that satisfy: $X_{i,j} = \frac{2i-1-H}{H}$, $Y_{i,j} = \frac{2j-1-W}{W}$. The value of each pixel is proportional to its distance from the origin in a specific direction and ranges from $[-1, 1]$ (Fig. 4). The expectation of the target coordinates $C = \{x, y\}$, could be derived as:

$$E(C(x, y)) = \{ \langle \hat{\mathcal{P}}, X \rangle, \langle \hat{\mathcal{P}}, Y \rangle \} \quad (1)$$

, where $\langle \cdot \rangle$ denotes the Frobenius inner product operation. As this operation is differentiable and fast to compute, it enables the model to learn from the ground truth (GT) coordinates directly. This meshgrid can be easily constructed and hard-coded into the layer before training, introducing no additional parameters.

Inserting the DSNT layer naively, however, does not guarantee good results. Two main issues are: 1) there is no constraint to penalize the presence of false-positive clutters - the appearance of normalized heatmaps could vary dramatically yet produce the same prediction; 2) limited supervision is provided to update the parameters of the whole network. To address both issues, a regularization term is appended to the loss function to penalize the spread of activations. This is achieved by controlling the variance of the modeled coordinate distribution $Var(C(x, y))$, computed as:

$$Var(C_v) = E(C_v)^2 - [E(C_v)]^2 = \langle \hat{\mathcal{P}}, C_v^2 \rangle - [\langle \hat{\mathcal{P}}, C_v \rangle]^2$$

, for each component $C_v \in x, y$. Given a variance controlling threshold φ , the overall variance could be computed as:

$$\mathcal{L}_{var} = \sum_{C_v \in x, y} (Var(C_v) - \varphi)^2 \quad (2)$$

The overall loss function is:

$$\mathcal{L}_{var} = \mathcal{L}_{coord} + \beta \mathcal{L}_{var} \quad (3)$$

, where β is a hyper-parameter to control the weight of the regularizer and \mathcal{L}_{coord} is the Euclidean distance between the ground truth and the prediction.

Backward projection Detecting facial landmarks from 2D images is an ill-posed problem. However, we can use the 2D results to infer the landmark locations in 3D utilizing the connection between the texture map and the polygon surface.

Specifically, we unwrap the 3D mesh on the whole texture map, then find the 3D polygon face that is the closet to the detected landmarks. As each face mesh consists of thousands of small surfaces, the area of each polygon surface is very small and could be regarded as a point. Therefore, the geometric center of this polygon is a good approximation of the target landmark location in 3D (Fig.5).

The derived 3D landmarks are then used to calculate PFL and utilized for subsequent surface-based analysis using dense surface modelling (DSM). The derived DSM, in turn, enables atypical facial morphology detection (e.g. lower-right in Fig. 4) and helps FAS-control discrimination.

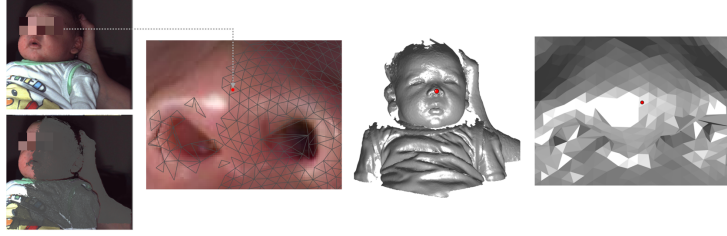


Fig. 5: Example of the 2D to 3D projection process. The nose tip is detected (red) on the 2D images. The 3D mesh is unwrap on the texture map (plotted as grey line). The closet 3D polygon surface to the landmark is found and its center is used to approximate the 3D position of the nose tip (plotted on the white 3D mesh).

4 Experiments

To demonstrate the effectiveness of the DSNT layer, we tested two comparative models. 1) Regression-based model (\mathcal{RB}). It uses a classical CNN model that consists of interlaced convolution (CONV) and down-sampling layers, followed by a fully-connected layer that predicts the target coordinates directly. 2) Synthetic heatmap matching. This method is tested with two different CNN models. The vanilla model (denoted as \mathcal{HM}_V) contains a down-sampling and up-sampling stream, the model branched out in the end into 20 prediction layers to predict the location heatmap of each landmark respectively. The second model follows the well-known work of [9], two vanilla CNNs are concatenated together into a hourglass-like model. This model (denoted as \mathcal{HM}_H) can capture information from different scales more freely which has proved to be helpful in landmark localization.

Implementation details The backbone of the tested CNN models contained four CONV block and down-sampling layers, while $\mathcal{HM}_V, \mathcal{HM}_H$ also have symmetrical up-sampling layers. A kernel of size 3×3 was used across the whole model. The numbers of feature channels for CONV layers are 8, 16, 32, and 64 for all remaining layers. Model training was done end-to-end via the Adam optimizer.

5 Results and Conclusion

Qualitative results of our model are shown in Fig. 6. The upper-right example in Fig. 6 is a particularly interesting case, as the subject is difficult to differentiate from the background due to the lighting conditions. However, the model was able to locate facial key-points accurately. Overall, the figure shows that our model is robust to appearance variations caused by extreme head pose (upper-left) facial expression (e.g. closed eyes in the upper middle and upper-right) and different ethnic background (lower-right and lower-middle).



Fig. 6: Predicted landmarks (red) on the test set. The images are resized to fit the page while the aspect ratio is kept to avoid distortion during visualization.

The localization accuracy was evaluated by calculating the Euclidean distance between the ground truth (GT) and the predictions in 2D (Tab. 1). It can be seen that our model has the best accuracy despite its smaller size. \mathcal{RB} has a large error, which may result from the fact that the GT landmark have large location variations therefore the model is confused in training. It is also interesting to see that \mathcal{HM}_V outperformed \mathcal{HM}_S in landmark detection by a small margin. This might be caused by overfitting as the \mathcal{HM}_S has a more complex network structure that requires more training data. A light-weight model is therefore memory-efficient and well-suited for our task.

The 2D detection results are propagated to the corresponding 3D meshes to obtain the 3D coordinates of each landmark. As children have fast growing trajectories, their faces can have very different sizes. To better evaluate the results, we adopt the popular normalized error metric:

$$norm_err = \frac{1}{N} \sum_i^N \frac{|d_i - g_i|_2}{|g_{re} - g_{le}|_2}, \quad (4)$$

while N is the number of evaluated landmarks, $|\cdot|_2$ is the Euclidean distance between two points ³, d_i and g_i are the predicted coordinates and the GT for the i_{th} landmark. $|g_{re} - g_{le}|_2$ is the inter-ocular distance: the Euclidean distance between the center of the two eyes. Note that we only use this metric in 3D as the annotations for both of the eyes do not co-exist in a single 2D image (see how the eyes are annotated in Fig.4). The accuracy results of our model is comparable to those reported on computer vision datasets [3], further proving the effectiveness of our model (Tab. 1). Table. 1 also reports the reliability of the automated generated PFL measurements. Our model has the smallest deviation from the GT (calculated based on manual annotations) while the \mathcal{RB} performed the worst. It should be pointed out the PFL derived from 3D meshes could be more reliable than from the current clinical method (placing a ruler against the subject face). The promotion of such a model might help automate and standardize this process.

Furthermore, the method allows more detailed morphology analysis of an individual against a control model (built based on [10]). Figure .4 (lower right) gives a preliminary but interesting example (red represents expansion, blue indicates compression). The difference image is created by calculating the point to point displacement from an individual to the control model. It can be seen clearly that the eyes are smaller (a strong FAS indicator), while the lip regions are similar to the normal (green represent neutral deformation). This first proves the complexity of FAS discrimination as there is no golden standard that is applicable to all. Further, it calls for the need to build more bespoke control models (e.g. based on ethnic background) which is a natural extension our current work.

Model	Param No.	2D Euc Dis	Norm-err 3D	PFL err (mm)
\mathcal{RB}	5.9	12.7 ± 17.3	45.8 ± 51.2	22.8 ± 19.6
\mathcal{HM}_V	1.8	0.76 ± 0.70	5.1 ± 4.4	1.4 ± 1.1
\mathcal{HM}_S	3.6	0.78 ± 0.72	5.3 ± 4.6	1.6 ± 1.3
<i>Ours</i>	0.6	0.67 ± 0.63	4.3 ± 3.9	1.2 ± 0.8

Table 1: Accuracy of different models. ‘Param No.’ is the number of parameter in millions. Euc Dic is the mean Euclidean distance between the GT and the prediction in 2D. Norm-err 3D is the percentage of normalized error distance calculated in 3D. ‘PFL err’ stands for the PFL measurement error in millimeters.

To conclude, we have presented a fully automatic landmark detection method to facilitate facial analysis and FAS detection of infants. As no human intervention is required, the method reduces the time burden and human effort greatly and facilitates analysis of larger populations in the future.

³ It is possible to calculate $|\cdot|_2$ using Geodesic distance, we choose Euclidean distance here to match the normalization factor—the inter-ocular distance, which is calculated using Euclidean distance.

References

1. B. M. Association *et al.*, “Alcohol and pregnancy: Preventing and managing fetal alcohol spectrum disorders. february 2016.”
2. T. S. Douglas and T. E. Mutsvangwa, “A review of facial image analysis for delineation of the facial phenotype associated with fetal alcohol syndrome,” *American Journal of Medical Genetics Part A*, vol. 152, no. 2, pp. 528–536, 2010.
3. Y. Wu and Q. Ji, “Facial landmark detection: A literature survey,” *International Journal of Computer Vision*, pp. 1–28, 2018.
4. C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652–660, 2017.
5. L. Yi, H. Su, X. Guo, and L. J. Guibas, “Syncspecnn: Synchronized spectral cnn for 3d shape segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2282–2290, 2017.
6. T. Le, G. Bui, and Y. Duan, “A multi-view recurrent neural network for 3d mesh segmentation,” *Computers & Graphics*, vol. 66, pp. 103–112, 2017.
7. J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” in *Conference on Fairness, Accountability and Transparency*, pp. 77–91, 2018.
8. A. Toshev and C. Szegedy, “Deeppose: Human pose estimation via deep neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1653–1660, 2014.
9. A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in *European Conference on Computer Vision*, pp. 483–499, Springer, 2016.
10. M. Suttie, T. Foroud, L. Wetherill, J. L. Jacobson, C. D. Molteno, E. M. Meintjes, H. E. Hoyme, N. Khaole, L. K. Robinson, E. P. Riley, *et al.*, “Facial dysmorphism across the fetal alcohol spectrum,” *Pediatrics*, vol. 131, no. 3, pp. e779–e788, 2013.
11. A. Nibali, Z. He, S. Morgan, and L. Prendergast, “Numerical coordinate regression with convolutional neural networks,” *arXiv preprint arXiv:1801.07372*, 2018.