Deep Learning-Based 3D Facial Landmark Localization: A Two-Step Approach with CNN and T-Net

Executive Summary

***Abstract*—This study introduces an automated 3D facial landmark annotation system using deep learning, specifically Convolutional Neural Networks (CNNs). Trained on stereophotogrammetry-acquired facial models, the model showcases potential applications in orthodontics, maxillofacial, and aesthetic surgery. By reducing manual annotation time, it enhances efficiency and precision in 3D anthropometric analysis, marking a significant stride in automating critical anatomical landmark identification. Furthermore, this study delves into stereophotogrammetry technology, highlighting its relevance in diverse fields. Despite requiring meticulous calibration, its advantages include minimal acquisition time and high accuracy. The conversion of 3D facial models into point cloud data facilitates versatile operations, with further exploration of point cloud processing methods, including PointNet. Emphasizing the crucial role of data pre-processing, ensuring a standardized and well-prepared dataset for effective model training. A two-stage deep learning approach, incorporating CNNs and T-Net architecture, achieves millimeter precision in facial landmark localization. The models undergo thorough analysis, considering variations in preprocessing techniques and their impact on performance. The Refined EsTi model emerges as robust and consistent, showcasing superior performance with adjusted metrics. These findings provide valuable insights for future developments in automated 3D facial landmark localization. The proposed methodology's versatility allows for comprehensive evaluations on diverse facial databases and landmarks, offering potential applications beyond the initial scope. This research represents a significant contribution to the field, demonstrating the potential of deep learning in automating critical anatomical landmark identification for medical and research purposes.**

Keywords— facial landmark localization, deep learning, pointnet, convolutional neural networks, stereophotogrammetry

I. Introduction

This study aims to automate the annotation of 3D facial landmarks leveraging deep learning methods such as CNNs and various architectures. The developed model, trained on a dataset of 200 stereophotogrammetry-acquired facial models, demonstrates potential applications in maxillofacial and orthodontics [1][2], and aesthetic surgery [3]. By reducing manual annotation time and ensuring a standardized approach, the tool enhances efficiency and precision in 3D anthropometric analysis. The work represents a significant step towards automating critical anatomical landmark identification, offering benefits in clinical and research settings [4], [5], [6], [7].

Recent studies have explored the integration of deep learning with 3D stereophotogrammetry for medical applications, such as infant head shape classification [8]. Despite challenges with small datasets, these approaches show promise in achieving high accuracy. Previous studies on Automatic landmark localization methods were investigated demonstrating effectiveness in predicting facial landmarks on 3D models. Creusot et al. [9] employ machine learning to annotate 14 facial landmarks on 3D models. They manually extract scalar and vector features from the meshes, including vector normal, neighboring point information, local volume, and principal curvatures. Offline training uses known landmark positions to learn statistical distributions, while online detection extracts feature and generates score maps to predict landmarks. Prediction error ranges from 2.5 to 10 mm compared to ground truth values. O'Sullivan and Zafeiriou [10] extended Convolutional Pose Machines (CPMs) for 3D facial landmark localization. They refined heatmaps using PointNet++ [11] architecture and a three-point strategy to mitigate outliers. Training involved minimizing mean squared error by using data augmentation on the BU-3DFE database. Their approach demonstrates the feasibility of extending 2D methods to 3D facial landmark localization. In the study by R. R. Paulsen et al [12], which adopted a multi-view approach for facial landmark identification, a localization error of 2.42mm was achieved on the BU-3DFE database [13]. While their study obtained a lower localization error compared to ours, it employed a more complex approach demanding significantly higher computational power. However, our study employed a simpler approach with lower computational requirements yet achieved comparable localization accuracy. It's noteworthy that in the study by R. R. Paulsen, the BU-3DFE database with its corresponding facial landmarks was used for training and evaluation, with all facial models cropped to contain only the face region. In contrast, our study utilized raw stereophotographs, introducing some limitations, which will be discussed in the next section. The aim of this study is to develop a deep learning model capable of localizing facial landmarks with the best localization accuracy.

II. Methods

A. Stereophotogrammetry

The fundamentals of stereophotogrammetry imaging technology are explored, emphasizing its significance in assessing facial morphology for diverse applications, including orthodontics, genetics, surgery, and forensic sciences [14]. Stereophotogrammetry, utilizing coordinated digital cameras and the stereoscopic principle, emerges as a promising approach for soft tissue analysis, offering advantages such as minimal acquisition time and high accuracy. Despite its benefits, stereophotogrammetry image acquisition requires meticulous calibration and post-processing, contributing to limitations such as equipment cost and fixed setups. The conversion of 3D facial models from stereophotogrammetry into point cloud data, representing the surface of physical objects as a collection of 3D data point,

allows for versatile operations and preprocessing methods such as ICP algorithm, and Morton code sorting.

B. *Neural Networks*

Artificial Neural Networks (ANNs) mimic the human brain's functioning and are a subset of Machine Learning (ML), particularly Deep Learning (DL). They automatically extract features from data through interconnected nodes called neurons. ANNs use activation functions to learn complex patterns, and training adjusts weights and biases to minimize errors. Regularization techniques prevent overfitting, and proper weight initialization ensures efficient training [15]. Convolutional Neural Networks (CNNs) specialize in visual data processing, using layers like convolutional, pooling, and fully connected layers. Convolutional layers extract features, pooling layers down sample data, and fully connected layers transform features into desired outputs. Max pooling selects essential features, making CNNs effective in tasks like object detection and classification [15].

PointNet [16] is a deep learning model designed for processing unordered point cloud data, making it versatile for 3D computer vision tasks. It's invariant to permutations, rotations, and translations, enhancing its segmentation performance. Using T-Net, PointNet transforms input data to a canonical form, inspired by Spatial Transformer Networks [17]. By sharing weights across its architecture, PointNet ensures consistent results regardless of point order. It's effective for tasks like classification and facial landmark localization [16] [18].

III. Experimental Protocols

The data was sourced from the LAFAS (Laboratory of Functional Anatomy of Stomatognathic system of Dipartimento di Scienze Biomediche per la Salute, Università degli Studi di Milano), containing 200 3D facial scans of healthy subjects. Each scan includes 50 facial landmarks annotated manually using a standardized protocol developed by Ferrario et al. [19]. These landmarks were labeled with precision using eyeliner on the face before acquisition. The 50 landmarks cover various facial regions, both midline and paired, as shown in figure 1.

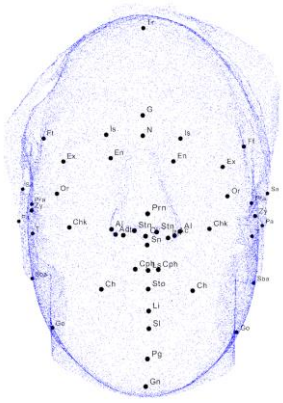

Fig. 1. 3D Facial with Identified Anatomical Facial Landmarks from C. Sforza et al . [20].

The 3D facial models were sampled to consist of 100,000 vertices using the Triangle Point Picking method [21], with landmarks of the models imported and stored. A preprocessing pipeline is established, demonstrated in figure 2, including alignments using the Iterative Closest Point (ICP) registration algorithm and cropping to focus only on the facial area. Resampling is done to ensure consistent number of vertices, followed by sorting based on spatial proximity using Morton code sorting or a reference point. Data augmentation techniques are applied, including rigid (rotation and scaling) and non-rigid (squeezing and stretching) methods. Vector normals of the point clouds are computed for each point, offering insights into surface orientation. Normalization is applied, and the point clouds along with their vector normals are stacked together for input to the model. The resulting data is stored in a multidimensional array for further analysis. These preprocessing steps aim to enhance the model's generalizability and accuracy in facial landmark localization, which will be evaluated in the results section of the study. Additionally, the dataset was split into training and testing sets, with 80% for training and 20% for testing, enabling robust evaluation of the model's performance.

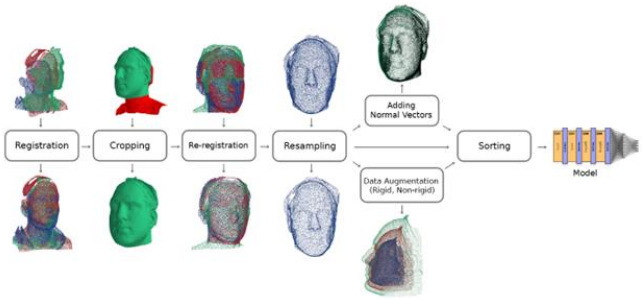

Fig. 2. *Preprocessing pipeline*

A two-stage DL model was utilized for facial landmark localization. The first stage, termed the Estimation stage, utilized pointwise CNNs for feature extraction, followed by fully connected layers for prediction. It consists of four convolutional blocks, each containing two pointwise layers and a MaxPooling layer. The second stage, the Refinement model, further enhances landmark localization accuracy by leveraging predicted coordinates from the Estimation stage. This involves isolating points within a predefined radius from the predicted landmarks and removing the rest. The points within the radius are stored and resampled to form sub-point clouds. These sub-point clouds undergo feature extraction through three convolutional blocks, followed by fully connected layers for prediction. MaxPooling is selectively applied to retain crucial features. The output layer produces predictions similar to the Estimation stage. Additionally, an EsTi model incorporating the T-Net architecture is introduced at the beginning of the Estimation stage to consider rotational and translational variations of faces in space. Figure 2 demonstrates the architecture of the EsTi model, where the T-Net architecture, in blue, produces a transformation matrix adjusted through the model's weights to transform point clouds into a canonical presentation. This transformed data is then fed into the Estimation stage, shown in orange. Models processed point clouds as multidimensional arrays, achieving predictions with millimeter precision. The Euclidean distance between

predicted and ground truth landmarks was utilized as the evaluation metric.

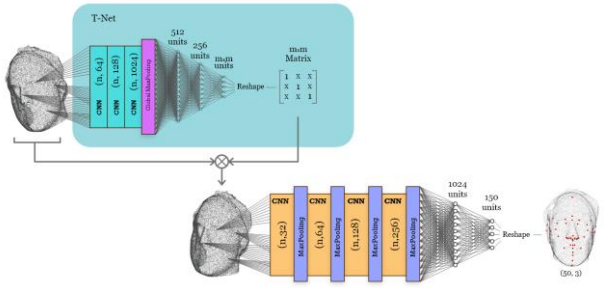

Fig. 3. *The architecture of the EsTi model*

In practical scenarios, predicted landmark coordinates may not precisely align with the surface of 3D facial models due to model complexity. To address this issue, two post-processing steps were introduced. One method selects the closest point in the point cloud to the predicted coordinate, while the other computes the centroid of the three closest points in sparse point clouds. These steps aim to address inaccuracies caused by model complexity and sparse data.

IV. Results and Discussion

Evaluation was conducted by comparing predicted coordinates with ground truth using Euclidean distance metrics, ensuring reproducible results with consistent initialization methods. Different preprocessing pipelines were explored, including baseline (Morton Code and reference point-based sorting), vector normals as features, data augmentation (rigid and non-rigid), and different sampling rates. The models, Refined Estimation, and Refined EsTi are measured based on Mean Euclidean Distance (MED) in both training and test sets, where the Euclidean distance of the predicted landmark and its ground truth are calculated and averaged across all the landmarks. Table 1 showcases the result for each of the preprocessing variations employed, for the models with the lowest localization error.

TABLE 1. *Localization error of the preprocessing variations*

PREPROCESSING	BEST MODEL	MED-train	MED-test
Baseline (Morton code sorting)	Refined EsTi	3.11mm	4.1mm
Baseline (Sorting based on the origin of the reference system)	Refined EsTi	2.8mm	3.9mm
Baseline ((Sorting based on a point above the forehead)	Refined EsTi	3.1mm	4.0mm
With Vector Normals	Refined Estimation	3.0mm	4.1mm
Data Augmentation (Rigid)	Refined Estimation & EsTi	4.0mm	4.2mm
Data Augmentation (non-rigid)	Refined Estimation	2.8mm	3.9mm
Resample (10,000 vertices)	Refined EsTi	3.4mm	4.6mm
Resample (20,000 vertices)	Refined Estimation & EsTi	3.1mm	4.3mm
Resample (30,000 vertices)	Refined EsTi	3.2mm	4.2mm

The refinement model showed the capability to improve predictions, adapting to error patterns from preceding models. Additionally, the incorporation of vector normals and data augmentation techniques like rigid and non-rigid augmentation were assessed, indicating varying impacts on model performance. Models trained on non-rigid augmentation and with vector normals showed improvement accuracy for the Refined Estimation model compared to the Refined EsTi model. Finally, different sampling rates were explored to balance computational complexity and localization accuracy, demonstrating the influence of point cloud density on model performance.

Among the evaluated models for 3D facial landmark localization, the Refined EsTi model, trained on point clouds indexed based on the origin of the reference system, and the Refined Estimation model, trained on non-rigid data augmentation, stood out with similar mean localization accuracy. However, a closer examination revealed distinctions in robustness. The EsTi model exhibited a more normal distribution of errors, frequent small errors, and lower residual standard deviation, indicating superior consistency and reliability. Adjusted metrics, excluding inconsistent landmarks, further emphasized the EsTi model's performance. These insights provide valuable guidance for future developments in this field. Figure 4 illustrates a 3D facial model, displaying its annotated ground truth landmarks depicted in green, alongside the predicted landmarks by the EsTi model highlighted in red, providing a visual comparison between the actual and predicted positions.

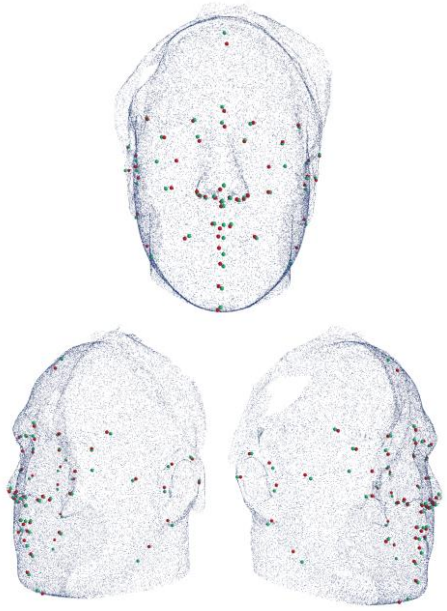

Fig. 4. *3D Facial Model with Predicted Landmarks (red) and Ground Truth Landmarks (green)*

V. Conclusion and Future Works

This study presents a deep learning-based method to automate 3D facial landmark annotation, beneficial for

medical applications. Leveraging techniques like stereophotogrammetry and CNNs, the models achieve millimeter precision. Preprocessing methods improve training, with the EsTi model showing consistency. To enable a comprehensive comparison with other studies, the proposed methodology can be evaluated on diverse facial databases and landmarks, including the BU-3DFE database. Furthermore, the feature extraction blocks of the model hold the potential to be trained on various 3D facial databases [13], [22], allowing them to capture facial features for different tasks. This versatility enables the model to perform tasks like facial detection and anatomical or non-anatomical facial landmark localization without the need for extensive training data.

Despite achieving promising results, limitations exist, including the small dataset size of 200 facial models and the absence of a dedicated validation set. Issues with the ICP registration algorithm and cropping process sensitivity to outliers are crucial, which can potentially impact the model's robustness. To address these challenges, future work could involve integrating more robust registration algorithms and exploring advanced cropping techniques. Additionally, incorporating RGB features based on point cloud positions and leveraging Graph Convolutional Neural Networks (GCNNs) could enhance model accuracy and reduce complexity. Evaluation on diverse facial databases and training feature extraction blocks on various datasets could further validate and extend the proposed methodology's capabilities.

REFERENCES

[1] C. Baserga *et al.*, "Efficacy of Autologous Fat Grafting in Restoring Facial Symmetry in Linear Morphea-Associated Lesions," *Symmetry*, vol. 12, no. 12, p. 2098, 2020.

[2] V. F. Ferrario, C. Sforza, J. H. Schmitz, and F. Santoro, "Three-dimensional facial morphometric assessment of soft tissue changes after orthognathic surgery," *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endodontology*, vol. 88, no. 5, pp. 549–556, 1999.

[3] J. B. Chang, K. H. Small, M. Choi, and N. S. Karp, "Three-Dimensional Surface Imaging in Plastic Surgery: Foundation, Practical Applications, and Beyond," *Plast. Reconstr. Surg.*, vol. 135, no. 5, p. 1295, May 2015, doi: 10.1097/PRS.0000000000001221.

[4] J. E. Allanson and T. R. P. Cole, "Sotos syndrome: Evolution of facial phenotype subjective and objective assessment," *Am. J. Med. Genet.*, vol. 65, no. 1, pp. 13–20, Oct. 1996, doi: 10.1002/(SICI)1096-8628(19961002)65:1<13::AID-AJMG2>3.0.CO;2-Z.

[5] S. Masnada *et al.*, "3D facial morphometry in Italian patients affected by Aicardi syndrome," *Am. J. Med. Genet. A.*, vol. 182, no. 10, pp. 2325–2332, 2020.

[6] C. Dolci *et al.*, "The face in marfan syndrome: A 3D quantitative approach for a better definition of dysmorphic features," *Clin. Anat.*, vol. 31, no. 3, pp. 380–386, 2018.

[7] J. E. Allanson and R. C. M. Hennekam, "Rubinstein-Taybi syndrome: Objective evaluation of craniofacial structure," *Am. J. Med. Genet.*, vol. 71, no. 4, pp. 414–419, Sep. 1997, doi: 10.1002/(SICI)1096-8628(19970905)71:4<414::AID-AJMG8>3.0.CO;2-T.

[8] G. de Jong *et al.*, "Combining deep learning with 3D stereophotogrammetry for craniosynostosis diagnosis," *Sci. Rep.*, vol. 10, no. 1, Art. no. 1, Sep. 2020, doi: 10.1038/s41598-020-72143-y.

[9] C. Creusot, N. Pears, and J. Austin, "A Machine-Learning Approach to Keypoint Detection and Landmarking on 3D Meshes," *Int. J. Comput. Vis.*, vol. 102, no. 1, pp. 146–179, Mar. 2013, doi: 10.1007/s11263-012-0605-9.

[10] E. O' Sullivan, "Extending Convolutional Pose Machines for Facial Landmark Localization in 3D Point Clouds," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0. Accessed: Jan. 16, 2024. [Online]. Available: https://openaccess.thecvf.com/content_ICCVW_2019/html/PreReg/Sullivan_Extending_Convolutional_Pose_Machines_for_Facial_Landmark_Localization_in_3D_ICCVW_2019_paper.html

[11] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Jan. 16, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/d8bf84be3800d12f74d8b05e9b89836f-Abstract.html

[12] R. R. Paulsen, K. A. Juhl, T. M. Haspang, T. Hansen, M. Ganz, and G. Einarsson, "Multi-view Consensus CNN for 3D Facial Landmark Placement," in *Computer Vision – ACCV 2018*, C. V. Jawahar, H. Li, G. Mori, and K. Schindler, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 706–719. doi: 10.1007/978-3-030-20887-5_44.

[13] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, *A 3D Facial Expression Database For Facial Behavior Research*, vol. 2006. 2006, p. 216. doi: 10.1109/FGR.2006.6.

[14] C. L. Heike, K. Upson, E. Stuhaug, and S. M. Weinberg, "3D digital stereophotogrammetry: a practical guide to facial image acquisition," *Head Face Med.*, vol. 6, no. 1, p. 18, Jul. 2010, doi: 10.1186/1746-160X-6-18.

[15] Ian Goodfellow and Yoshua Bengio and Aaron Courville, *Deep Learning*. MIT Press.

[16] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation.” arXiv, Apr. 10, 2017. doi: 10.48550/arXiv.1612.00593.

[17] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial Transformer Networks.” arXiv, Feb. 04, 2016. doi: 10.48550/arXiv.1506.02025.

[18] F. Zhang, J. Fang, B. Wah, and P. Torr, “Deep FusionNet for Point Cloud Semantic Segmentation,” in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 644–663. doi: 10.1007/978-3-030-58586-0_38.

[19] V. F. Ferrario, C. Sforza, G. Serrao, V. Ciusa, and C. Dellavia, “Growth and aging of facial soft tissues: A computerized three-dimensional mesh diagram analysis,” *Clin. Anat.*, vol. 16, no. 5, pp. 420–433, 2003, doi: 10.1002/ca.10154.

[20] C. Sforza, C. Dellavia, M. De Menezes, R. Rosati, and V. F. Ferrario, “Three-dimensional facial morphometry: from anthropometry to digital morphology,” in *Handbook of Anthropometry: Physical Measures of Human Form in Health and Disease*, Springer, 2012, pp. 611–624.

[21] E. W. Weisstein, “Triangle Point Picking.” Accessed: Feb. 05, 2024. [Online]. Available: https://mathworld.wolfram.com/

[22] A. Savran *et al.*, “Bosphorus Database for 3D Face Analysis,” in *Biometrics and Identity Management*, B. Schouten, N. C. Juul, A. Drygajlo, and M. Tistarelli, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2008, pp. 47–56. doi: 10.1007/978-3-540-89991-4_6.