Gender Classification via Graph Convolutional Networks on 3D Facial Models

Giorgio Blandano
University of Milan
Department of Computer Science
giorgio.blandano@studenti.unimi.it

Jacopo Burger
University of Milan
Department of Computer Science
jacopo.burger@studenti.unimi.it

Claudia Dolci
University of Milan
LAFAS - Department of Biomedical Sciences for Health
claudia.dolci@unimi.it

Giuseppe M. Facchi
University of Milan
Department of Computer Science
giuseppe.facchi@unimi.it

Federico Pedersini
University of Milan
Department of Computer Science
federico.pedersini@unimi.it

Chiarella Sforza
University of Milan
LAFAS - Department of Biomedical Sciences for Health
chiarella.sforza@unimi.it

Gianluca M. Tartaglia
University of Milan
Department of Biomedical, Surgical and Dental Sciences
Ospedale Maggiore Policlinico
UOC Maxillo-Facial Surgery and Dentistry Fondazione IRCCS Cà Granda
gianluca.tartaglia@unimi.it

Annalisa Cappella
University of Milan
LAFAS - Department of Biomedical Sciences for Health
IRCCS Policlinico San Donato
U.O. Laboratorio di Morfologia Umana Applicata
annalisa.cappella@unimi.it

ABSTRACT

The automatic classification of human gender and other demographic attributes such as age and ethnicity is gaining significant attention. These attributes provide rich information with applications in personalization, behavior analysis, consumer research, digital forensics, security, human-computer interaction, and mobile applications. In the literature, the face is a commonly used feature for gender classification. In this paper, we follow this attitude but referring to 3D face data that offer advantages in terms of capturing spatial information and reducing sensitivity to ethnicity and acquisition conditions. In particular, we address gender classification using RGB-D data, which is structured as graphs and processed using a Graph Convolutional Neural Network (GCNN). Experiments conducted on the BP4D+ dataset demonstrate the effectiveness of this approach.

CCS CONCEPTS

• **Computing methodologies** → **Graph Convolutional Neural networks**.

SAC '24, April 8–12, 2024, Avila, Spain

ACM ISBN 979-8-4007-0243-3/24/04…$15.00
https://doi.org/10.1145/3605098.3636039

KEYWORDS

Graph Convolutional Neural networks, RGB-D faces, Gender recognition, BP4D+ dataset

ACM Reference Format:
Giorgio Blandano, Jacopo Burger, Claudia Dolci, Giuseppe M. Facchi, Federico Pedersini, Chiarella Sforza, Gianluca M. Tartaglia, and Annalisa Cappella. 2024. Gender Classification via Graph Convolutional Networks on 3D Facial Models. In *Proceedings of ACM SAC Conference (SAC'24).* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3605098.3636039

1 INTRODUCTION

Automatic classification of human gender, and other demographic attributes such as age and ethnicity are gaining significant attention due to the rich and distinct information these attributes provide [33]. Indeed, numerous domains, including personalization and recommender systems, behavior analysis, consumer research, digital forensics, security and biometrics, human-computer interaction, and mobile applications, stand to improve their performance by having access to user gender information [22].

According to the literature, the approaches for gender detection based on data derived from the human body can be classified as appearance-based and non-appearance-based. The appearance-based approaches can be further categorized in using static body features (face, hands, fingernails, body shape), dynamic body features (gesture, motion, gait, voice), and clothing features (clothing, footwear). The non-appearance-based approaches use biometric features (fingerprint, iris, ear, skin colour), bio-signals (DNA, EEG, ECG), and social information (blog, email, handwriting) for gender detection.

Beyond doubts, the face is the most frequently employed feature in gender classification [2, 10, 17, 36, 37], as it is easy to capture and serves other purposes such as identity [9, 11] and expression recognition [4, 20]. According to the field of application, face images can be characterized by RGB-images acquired either in the wild [17] or in more controlled conditions [13]. Certainly, when it comes to evaluating anthropometric facial features, RGB-D data would be the most suitable choice [40]. Indeed, three-dimensional data captures the spatial information of the face, providing a richer representation of facial characteristics. By incorporating depth information, 3D data can offer a more comprehensive view of facial features, making it better suited for multiple classification tasks, like gender classification. Furthermore, with the growing availability and popularity of devices designed for capturing RGB-D data, it is evident that such data could become a practical and significant option in many application, including automatic gender recognition.

The benefit of 3D face representation lies in its emphasis on geometric attributes rather than appearance-based characteristics. This shift diminishes sensitivity to both ethnic differences, including facial features such as skin tone, and hairstyle, as well as acquisition conditions such as lighting, camera angles, and image quality, thereby enhancing generalization across diverse populations and acquisition scenarios.

Indeed, 3D data poses distinct challenges. Primarily, it is susceptible to changes in facial expressions that alter facial structure. Furthermore, the current lack of a sizable dataset containing RGB-D faces gives rise to legitimate concerns when considering the use of deep learning methods, notably requiring large volume of labeled data in order to avoid overfitting and bias in the model's prediction.

In this paper, we address the gender classification problem using RGB-D data, and to achieve this, we suggest employing a Graph Neural Network (GNN), due to its established capacity for robust generalization both on 2D and 3D data [25, 30]. Specifically, we implement a Graph Convolutional Neural Network (GCNN) [38], being highly adept at capturing both the local and global geometric features present in 3D facial data. Indeed, GCNNs exhibit a remarkable capability to gather information from nearby facial landmarks, facilitating a comprehensive understanding of the intricate nuances within facial structures. This competence proves to be pivotal, particularly in tasks like gender classification, where the discernment of subtle geometric patterns indicative of gender plays a crucial role.

The experiments involve conducting comparisons using the BP4D+ dataset [39], which comprises recordings from 140 participants with a large range of attributes, including gender, age, facial expressions, and ethnic variations. These experiments with the state-of-the-art methods in this field, allow to show the effectiveness of our approach.

2 RELATED WORKS

To our knowledge, the latest survey on gender recognition date back 2016 [22], thus not including deep learning based solutions. Traditionally, methods required a feature extraction module to extract some spatial or textural feature (e.g. LBP, wavelets, ...), followed by a classification module (e.g. support vector machines, linear discriminant analysis, ...) [3, 29].

Since then, Convolutional Neural Networks (CNNs) have emerged as the leading approach in numerous computer vision tasks, including gender recognition, often tackled together with other demographic tasks (e.g. age, ethnicity). Levi and Hassner were among the first proposing a simple CNN architecture with 5 layers, to perform age and gender prediction [19]. Besides, the effectiveness of employing pre-trained CNNs like AlexNet or VGG has been exemplified in numerous studies, as evidenced by papers such as [18, 24, 33].

Advancements have been achieved integrating attention mechanisms into the feed-forward models, enabling these models to identify the most informative and reliable facial components for the specific task at hand, as demonstrated in [27]. In the same vein, Abdolrashidi el al. [1] proposed the ensemble of attentional and residual convolutional networks. The effectiveness of attention mechanisms has been amplified with the advent of transformers [34], which have been consistently gaining prominence and importance in various tasks and benchmarks. Gender recognition is no exception, and Kuprashevich and Tolstykh [17] introduced a transformer model, namely MiVOLO, for age and gender estimation, establishing them as the current state-of-the-art approach.

To the best of our knowledge, no model working on 3D data has been proposed yet.

3 LANDMARK-BASED GRAPH STRUCTURE

When working with a 3D facial mesh, various approaches could be employed for classification tasks. Direct manipulation of facial point clouds is one option, as demonstrated in techniques like PointNet [30]. Another method could revolve around working with the graph structure that defines the mesh itself.

Let $M = (P, C)$ denotes a mesh, with $P = \{p_i\}_{i=1}^{N} \subset E^3$ representing the points in the 3D euclidean space and C the connections between them. Due to the large number of points in a mesh, we adopt a method inspired by anthropometry in medicine that leverages a set of facial landmarks to capture key features of the human face [7, 8].

This way, we define a landmarks-based graph $G = (V, E)$, where $V = \{v_i\}_{i=1}^{L} \subset P$, with $L \ll N$, is a set of extracted facial landmarks and E the edges computed via k-NN(v_i), for each $i \in [1..L]$. The rationale behind this approach is to focus on few significant points, and characterize them with a robust discriminative representation (cfr. Sec. 3.1) such as surface curvatures or geometric relations such as distances.

There exists a significant literature about 3D Facial landmark detection [14], with methods falling into two main categories: those based on 3D geometric information [21, 23], and those relying on statistical learned models [15, 32].

Here, we adopt a statistical learned model, more precisely the MVLM proposed by Rasmus et al. [26]. Briefly, MVLM exploits the 3D face mesh to render several views from different view points and uses a CNN based model to estimate the 2D location of each landmark from each view point. These estimates are then combined using a LSQ and RANSAC approach to have a robust and reliable estimate of the 3D location of $L = 84$ landmarks. The MVLM model has been deployed in several pre-trained versions with a combination of different datasets and rendering methods, these include RGB renderings as well as depth and geometry ones. In our work

we refer to the version trained using geometry+depth image channels and with the BU-3DFE dataset which contains 100 subjects (56 female, 44 male), that gained an average error of 2.42 mm on localizing 84 3D landmarks (Figure 1).

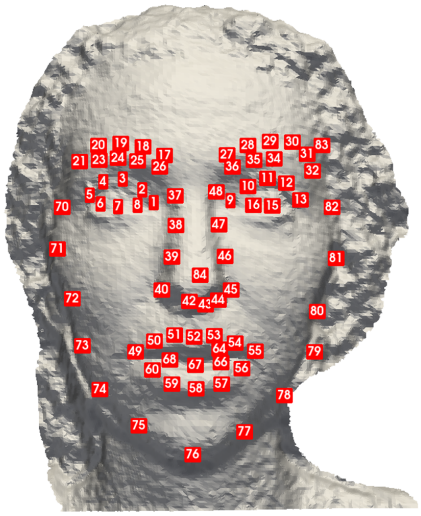

Figure 1: Example of a BP4D+ subject triangular mesh, with the 84 extracted landmarks (not all visible)

As an example, in Figure 2 three different views of a graph with 84 landmarks are reported.

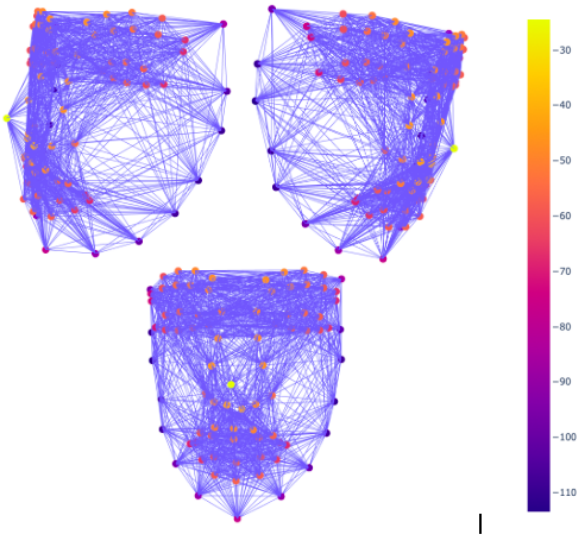

Figure 2: Different views of a graph obtained defining a node for each of the 84 landmarks and linking each node to its k nearest neighbours, with $k = 32$. The graduated scale provides information about z-coordinate of each landmark.

3.1 Node-level Feature Representation

In this section, a joint multimodal embedding space that includes both local geometric features and simple empirical statistics is proposed and motivated.

As for geometric features we simply use the 3D coordinates of each landmark $v_i \in V$, i.e. $\text{pos}(v_i) = (x_i, y_i, z_i)$ and for each pair of distinct landmarks $v_i, v_j \in V$ we compute the geodesic distance from v_i to v_j. The computation of approximate shortest (i.e. geodesic) paths on a triangle mesh is a common operation in many computer graphics applications [5]. In particular, here we implemented a method based on the heat equation [6], taking the geodesic distances $\text{geo}(v_i)$ between a node $v_i \in V$ and all points in its neighborhood $\mathcal{N}(v_i)$.

Furthermore, following [28], we introduce a feature called Fast Point Feature Histograms (FPFH), a method for representing feature points in 3D point cloud data, which can be used to accurately label the points based on the type of surface they are lying on. This approach aims to capture the geometric relationships between each point in a 3D point cloud and its neighboring points, going beyond surface normals and curvature estimations to characterize the mean curvature surrounding a specific point. The proposed feature representation is based on a multi-dimensional histogram that characterizes the local geometry around a query point. A key property of this representation is its invariance to pose (i.e. 3D rotations and translations) and sampling density, and it can cope well with noisy sensor data.

To define the feature space computational model, the authors of [28] introduce the following elements. For each pair of nodes v_i and v_j, we build a fixed reference frame, consisting of the three unit vectors s, t, z defined as follows:

$$
\begin{aligned}
&s \text{ is the surface normal } n_i \text{ at } v_i \\
&t = s \times \frac{v_i - v_j}{\|v_i - v_j\|_2} \\
&z = s \times t,
\end{aligned}
$$

where $\times$ denotes the cross product of two vectors.

Using the defined reference frame, the difference between the two normals n_i and n_j can be expressed as a set of angular features:

$$
\begin{aligned}
\alpha &= t \cdot n_j \\
\phi &= s \cdot \frac{v_i - v_j}{\|v_i - v_j\|_2} \\
\theta &= \arctan(z \cdot n_j, s \cdot n_j),
\end{aligned}
$$

where $\cdot$ denotes the scalar product of two vectors. The attributes α and θ represent n_j as an azimuthal angle and the cosine of a polar angle, respectively, while ϕ represents the direction of the translation from v_i to v_j.

In the model proposed in [28], the 3D feature distribution of points sampled from the 3D face surface is represented by histograms. In particular, for a point v_i each feature of the tuple (α, ϕ, θ) is mapped onto exactly one bin a, b, c of the histograms $H_{\alpha,a}(v_i)$, $H_{\phi,b}(v_i)$ and $H_{\theta,c}(v_i)$, respectively, where $a, b, c \in [1..11]$, thus providing three 11-dimensional histograms for each tuple. The final 33-dimensional histogram SPFH (Simple Point Feature Histogram) for each vertex v_i is achieved by horizontally concatenating

the three given histograms, i.e.

$$\text{SPFH}(v_i) = [H_\alpha(v_i), H_\phi(v_i), H_\theta(v_i)].$$

.

The final FPFH descriptor representing each point in a face is computed by collecting the local information of each landmark point $v_i \in V$ on the face:

$$\text{FPFH}(v_i) = \text{SPFH}(v_i) + \frac{1}{|\mathcal{N}(v_i)|} \sum_{v_j \in \mathcal{N}(v_i)} \frac{1}{\delta_{ij}} \text{SPFH}(v_j) \quad (1)$$

being δ_{ij} the distance between the query node v_i and its neighbor v_j.

In our approach the node characterization is obtained by horizontally concatenating the 3 vectors, i.e. sets of features defined above, that are point coordinates (3-dim), geodesic distances (84-dim) and histograms (33-dim), yielding the comprehensive 120-dimensional vector:

$$h(v_i) = [\text{pos}(v_i), \text{FPFH}(v_i), \text{geo}(v_i)] \quad (2)$$

The histograms for each node, interpreted as embeddings in a 120-dimensional space, enables the use of a GNN model introduced in the next section, where the nodes of the graph model correspond to landmarks of the face.

4 GENDER-GCNN RECOGNITION METHOD

Graph Neural Networks (GNNs) employ multiple layers to process node representations. In a nutshell, under the assumption of a static graph and a single input feature vector per node, at each layer indexed by $l \geq 0$ (with $l = 0$ representing the input layer), GNNs calculate a node representation by gathering information from its neighborhood through an aggregator. Stacking r layers in a GNN allows a node's r-hop neighborhood to affect its representation.

Since we are dealing with a prediction problem for the entire graph (graph-level task), we need to consider the relationships between all of the nodes in the graph. In our scenarios, the graph structure is explicitly induced by the landmark-based graph described in the previous section, where the embeddings initially assigned to each node v_i correspond to those calculated in Eq. (2). The edges in the graph are undirected, with no features, and are built linking each node to its k nearest neighbours, with $k = 32$.

Formally, let $h^l(v)$ be the embedding of node v at layer l, being $h^0(v)$ defined in Eq. (2). A nonlinear aggregation mechanism is employed for the evolution of $h^l(v)$, taking into account its embedding at the previous layer, $h^{l-1}(v)$, as well as those of its neighbors in $\mathcal{N}(v)$. This way, in each message-passing iteration of a GNN, the embedding $h^l(v)$ for each node $v \in V$ is updated based on information aggregated from v's neighbors, that is

$$\begin{aligned} m^l(v) &= \text{MSG}^l\left(\left\{h^{l-1}(u), \forall u \in \mathcal{N}(v)\right\}\right) \\ h^l(v) &= \text{AGG}^l\left(h^{l-1}(v), m^l(v)\right) \end{aligned} \quad (3)$$

where, MSG^l and AGG^l are arbitrary differentiable functions representing message computation and aggregation, respectively. While MSG^l employs a Multi-Layer Perceptron (MLP) network, AGG^l can implement various aggregators, such as graph convolution [16], attention mechanisms [35], or pooling [12].

In this work Eq. 3 is implemented in `Gender-GCNN Layer` according to these two phases:

(1) Message computation: Each node receives messages from all its 32 neighbors, and each message is computed using a shared-weight MLP, i.e.

$$m_{ij}^l = \text{MLP}^l\left(\left[\delta_{ij}^{l-1}(v_i), s^{l-1}(v_i)\right], v_j \in \mathcal{N}(v_i)\right), \quad l > 0,$$

where $\delta_{ij}^l(v_i) = h_{:3}^l(v_i) - h_{:3}^l(v_j)$ and $s^l(v_i) = h_{3:}^l(v_i)$, being $h_{:3}$ the first 3 entries of vector h and $h_{3:}$ the complementary set.

(2) Node Features Update: The node features are aggregated using the MAX operator, i.e.

$$h^l(v_i) = \text{MAX}\left(m_{ij}^l, \forall v_j \in \mathcal{N}(v_i)\right), \quad l > 0,$$

where the MAX function is used to combine the information from all of a node's neighbors messages into a single vector. This vector is then used to update the node's own features, so that the node can learn to represent its neighbors and itself in a more informative way.

Our Gender-GCNN final network model (see Figure 3) consists of four `Gender-GCNN Layers`. After a first convolutional layer that preserves the feature dimensionality, the network progressively reduces the feature dimensionality from 120 to 64, then to 32, and finally to 16.

Following the four `Gender-GCNN Layers`, the information of all nodes is aggregated in a single super-node, creating a representation of the entire graph via a global max-pooling approach, that is each feature of the super-node is determined by taking the maximum value among the corresponding features of all the nodes in the graph. By following this approach, it is expected that the super-node will contain all the relevant information and fully characterize the associated graph. Subsequently, the features of the super-node are passed to the final classifier, which is implemented as a fully-connected (FC) layer. The FC layer includes a softmax layer at the end, which normalizes the sum of the predicted values for the two target classes to 1. By applying this normalization, the output can be interpreted as the model's predicted probability of the input being associated with each class.

5 EXPERIMENTAL RESULTS

In this section, we describe the experimental setup and methods we used to compare our proposed Gender-GCNN model, which works on 3D data, to other deep learning models designed for gender classification. In the following we detail the dataset used, the models referred for comparisons, the experimental protocol, and the obtained results.

5.1 The dataset

The multimodal spontaneous emotion (MMSE) dataset [39] also referred to as BP4D+ consists of 1400 multimodal recordings from 140 participants varying for gender, age, and racial ancestries. Data includes for each recording a 3D dynamic facial models (each model has 30-50k vertices), two 2D texture videos acquired at 25fps, with resolution 1040×1392 pixels (acquired using the Di3D dynamic

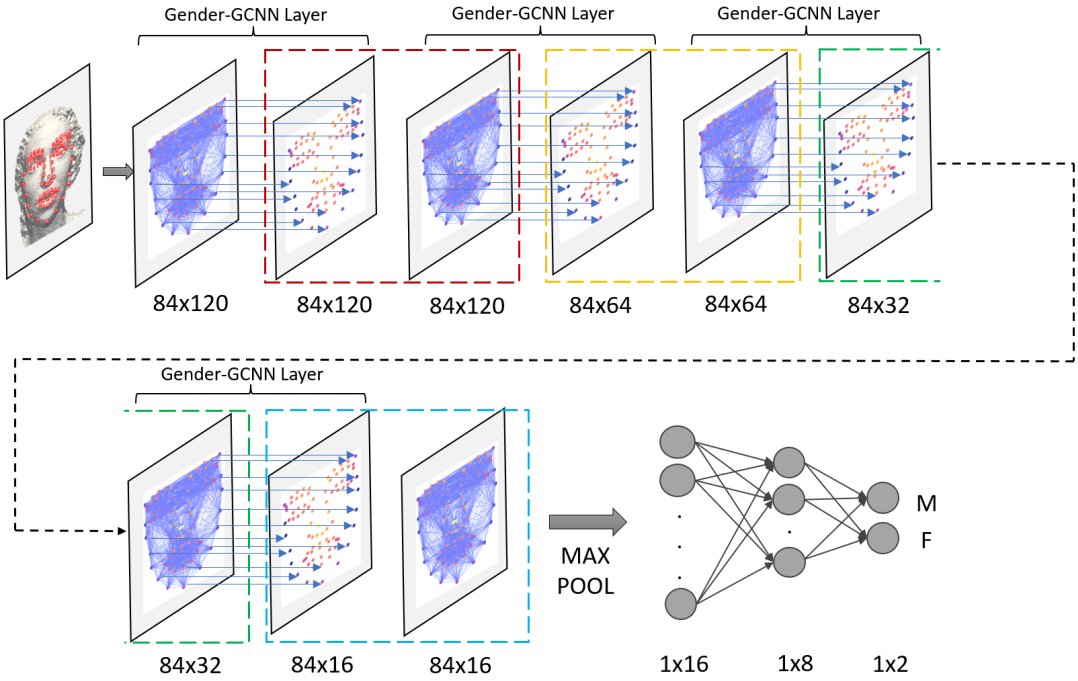

Figure 3: Our Gender-GCNN model. It consists of 4 `Gender-GCNN Layer`, where each layer shares the same MLP weights. After a max-pooling operation, a MLP performs the final classification. MLPs are represented as light blue arrows.

Figure 4: An example of a BP4D+ record. From left to right, a frame's 2D RGB texture, plain mesh and RGB mesh.

imaging system), a thermal video (acquired at 25 fps with resolution 640 × 480 pixels). An example of a record from the dataset is represented in Figure 4.

Furthermore, the Biopac MP150 data acquisition system is used to collect physiological signals, including blood pressure, respiration frequency, and EDA. Participants took part to 10 different tasks conceived to elicit different emotions. Each frame is manually labelled with its corresponding Action Unit (AU) vector, which delineates the activation status of each AU: 0 signifies inactivity, while 1 indicates activity.

In this study, only the 2D textures, 3D meshes and the AU annotations of 136 people have been utilized, being the publicly available ones.

5.2 Comparison models

To evaluate the performance of our proposed 3D-based method, we compare it to two state-of-the-art 2D gender recognition models.

The first model we consider is DeepFace [31], which is a facial recognition system developed by Facebook (now Meta Platforms, Inc.) that uses deep learning algorithms to identify and verify faces in photos and videos. It gained attention for its ability to achieve high accuracy in facial recognition tasks, including gender estimation and emotion analysis.

The second is MiVOLO (Multi Input VOLO) [17], which is a recent work that proposes a straightforward approach for age and gender estimation using the latest vision transformer. MiVOLO provides several pre-trained models, each with a different combination of input (face, body) and output (age, gender). The one used in this work requires only the face as input and provides age and gender as output.

5.3 Experiments

As previously mentioned, we carried out the experiments utilizing the BP4D+ dataset. As an initial study, we evaluated the methods under favorable conditions, reducing the complexity introduced by variations in facial expressions. To achieve this, we focused on data from a single task out of the 10 included in BP4D+. Specifically, we utilized data from Task 1, which is characterized by relaxed reactions and happiness responses.

Within this scenario, we constructed multiple datasets, gradually relaxing the constraint over the neutrality of facial expressions. More specifically, a perfect neutral frame can be described as one

where specialized psychologists carefully evaluate the action units, ensuring that their cumulative total equals 0. However, this request is very hard to be verified in the reality, so we relaxed this criterion at different levels, selecting, when possible, 10 frames for each person, where the action units sum up to less than 3, 4, or 5 respectively.

Let V_s be the video of the s-th subject in the dataset and F_s be a collection of its frames selected according to the following rule:

$$F_s = \{x \mid x \in V_s\,,\ AU(x) < \theta\}, \quad |F_s| \leq 10,$$

where $\theta \in \{3, 4, 5\}$ and $AU(x)$ is the sum of the AUs of the frame x. The final dataset for a given level θ is then given by:

$$D_\theta = \bigcup_{s=1}^{S} F_s$$

where S is the number of subjects in the original dataset.

Additionally, we generated two more datasets with even less control over the neutrality of expressions, not relying on psychologist labeling. In one case, we leveraged the automatic expression classification provided by DeepFace, resulting in the dataset D_{DF}. In the other instance, we performed random sampling of frames from the entire video collection without considering specific facial expressions, yielding D_{rand}.

The D_{DF} dataset collects up to 10 frames for each subject by identifying frames where the subject is making a neutral expression, with a neutral score greater than a threshold we fixed to 95%.

To obtain D_{rand}, we performed a random sampling of 10 frames from each video without any control on the represented faces. This approach resulted in the inclusion of frames exhibiting a variety of facial expressions and occasional occlusions caused by the movements of various body parts. For instance, some frames might feature an arm obstructing the camera view. These variations represent potential challenges in the classification processes. It is also worth noting that dataset D_{rand} includes a wider range of facial expressions, which makes it more challenging to train a model on. However, it also makes the model more generalizable to new data, since it is less likely to have overfit to the training data.

The obtained datasets are characterized by low cardinality being the limited amount of data particularly strong when referring to the very selective dataset D_3, incorporating 94 out of 136 subjects. This is due to the fact that several videos exhibit lower or any frames with total AUs score lower than 3. Similarly, the datasets D_4 and D_5, incorporate 111 and 122 subject, respectively. Finally, D_{DF} and D_{rand}, comprising 135 and 136 subjects respectively. Furthermore, it's worth noting that with the exception of D_{rand}, which contains precisely 10 frames per subject, the other datasets have, on average, 9 frames per subject meeting the specified criteria. This results in the cardinalities detailed in Table 2, in the first row.

The datasets generated were used for the purpose of evaluating the performance of the Gender-GCNN model, the Transformer-based system MiVOLO, and the DeepFace models. Data is used at frame level, meaning that 2D images are used to test the MiVOLO and DeepFace models, and the 3D models corresponding to the chosen frames are used to evaluate the Gender-GCNN model. For our experiments, we employed a pre-trained MiVOLO model provided by the authors, which was originally trained on the IMDB dataset. As for DeepFace, experiments were carried out using a model based on a pre-trained VGG-Face model, where both IMDB and WIKI datasets had been used for both training and testing phases.

To train the Gender-GCNN model, we used leave-one-subject-out cross-validation (LOSOCV) to address the problem of limited data in each dataset. LOSOCV comes with several advantages. It helps to ensure that the model is robust by evaluating its performance on a wide variety of subjects in the dataset. This allows us to make the most of the available data and get accurate performance estimates. This technique also aids in mitigating biases that may exist in the dataset, reducing the impact of subject-specific idiosyncrasies. Moreover, we enhanced the robustness of our metrics by calculating both mean accuracy and standard deviation through multiple trials (specifically, 10 trials have been setup) of LOSOCV. A summary of the hyperparameters used for training procedure is listed in Table 1. To address the imbalance in the dataset classes, we also implemented an oversampling technique. This technique involved duplicating samples from a subset of the training data until a perfect balance between male and female classes was achieved.

Table 1: Training Procedure Hyperparameters

Hyperparameter	Value
Learning Rate	0.001
Batch Size	64
Number of Epochs	100
Optimizer	Adam
Loss Function	Binary Cross Entropy

Table 2 presents the performance results achieved by these models when tested on the datasets described earlier.

Despite being trained on datasets with significantly fewer frames than the MiVOLO and DeepFace models, the Gender-GCNN model achieves superior gender recognition performance, with a mean accuracy of 90.41% and a minimal standard deviation. This remarkable ability is due to several inherent strengths of our model. One of the key strengths of the Gender-GCNN model is its ability to capture subtle geometric features that are invariant to ethnicity, appearance, and facial expression. Unlike traditional models, which can struggle with variations in these factors, GCNNs are able to learn hierarchical features that transcend these differences. This adaptability to geometric features allows the Gender-GCNN model to generalize effectively and robustly, even when trained on limited data. Another strength of the Gender-GCNN model is its feature engineering process, which leverages 3D data to compute FPFH and geodesic distances. These features provide a more accurate representation of the facial structure than 2D images, which inherently lose depth information. 3D data also retains spatial relationships and surface curvatures, leading to a richer and more realistic portrayal of the face.

6 CONCLUSIONS

The automatic classification of human gender and other demographic attributes, has gained significant attention in various domains. 3D data offers several advantages for gender classification,

Table 2: Summary of the experiments conducted on the different datasets, with the number of frames for each dataset in parentheses.

Model	$D_3(878)$	$D_4(1041)$	$D_5(1207)$	$D_{\text{DF}}(1221)$	$D_{\text{rand}}(1360)$	Mean
MiVOLO D1	87.13%	89.15%	88.48%	87.04%	86.91%	87.74%
DeepFace	68.11%	69.26%	66.44%	66.73%	68.01%	67.71%
Ours (Gender-GCNN)	90.07% ± 0.18%	90.71% ± 0.35%	90.97% ± 0.89%	91.29% ± 0.52%	90.37% ± 0.49%	90.68% ± 0.48%

including increased robustness across diverse populations and acquisition scenarios, as well as better representation of facial characteristics. However, challenges remain in terms of data availability and changes in facial expressions, which require further investigation. The combination of our Gender-GCNN model's inherent strengths in capturing geometric invariances with the strategic feature engineering process focusing on geometrical and morphological cues leads to enhanced discriminative power. This two-fold approach empowers our model to excel in gender recognition, even in scenarios with limited training data cardinality, offering a more refined and accurate understanding of facial gender characteristics.

Additional work is needed to evaluate the robustness, generalization, and applicability of our model to a wider range of 3D faces datasets, including those from diverse populations, acquisition scenarios, and real-world settings. Finally, building upon the promising performance observed in this context, it would be valuable to explore the potential application of this approach in other domains that face the challenge of low data cardinality. For instance, the medical field, where the objective is to diagnose rare diseases based on facial characteristics, could benefit from such investigations.

REFERENCES

[1] Amirali Abdolrashidi, Mehdi Minaei, Elham Azimi, and Shervin Minaee. 2020. Age and gender prediction from face images using attentional convolutional network. *arXiv preprint arXiv:2010.03791* (2020).

[2] Norah AlShaye, Lamia AlMoajil, and M Abdullah-Al-Wadud. 2022. A Gender Recognition System Based on Facial Image. In *2022 3rd International Conference on Artificial Intelligence, Robotics and Control (AIRC)*. IEEE, 21–25.

[3] Amjath Fareeth Basha and Gul Shaira Banu Jahangeer. 2012. Face gender image classification using various wavelet transform and support vector machine with various kernels. *International Journal of Computer Science Issues (IJCSI)* 9, 6 (2012), 150.

[4] Sathya Bursic, Giuseppe Boccignone, Alfio Ferrara, Alessandro D'Amelio, and Raffaella Lanzarotti. 2020. Improving the accuracy of automatic facial expression recognition in speaking subjects with deep learning. *Applied Sciences* 10, 11 (2020), 4002.

[5] Keenan Crane, Marco Livesu, Enrico Puppo, and Yipeng Qin. 2020. A survey of algorithms for geodesic paths and distances. *arXiv preprint arXiv:2007.10430* (2020).

[6] Keenan Crane, Clarisse Weischedel, and Max Wardetzky. 2017. The heat method for distance computation. *Commun. ACM* 60, 11 (2017), 90–99.

[7] Claudia Dolci, Valentina Pucciarelli, Marina Codari, Susan Marelli, Giuliana Trifiro, Alessandro Pini, Chiarella Sforza, et al. 2016. 3D morphometric evaluation of craniofacial features in adult subjects with Marfan syndrome. In *Proceedings of the 7th International Conference on 3D Body Scanning Technologies, Lugano, Switzerland, Hometrica Consulting*. 98–104.

[8] Claudia Dolci, Valeria A Sansone, Daniele Gibelli, Annalisa Cappella, and Chiarella Sforza. 2021. Distinctive facial features in Andersen–Tawil syndrome: A three-dimensional stereophotogrammetric analysis. *American Journal of Medical Genetics Part A* 185, 3 (2021), 781–789.

[9] Hang Du, Hailin Shi, Dan Zeng, Xiao-Ping Zhang, and Tao Mei. 2022. The elements of end-to-end deep face recognition: A survey of recent advances. *ACM Computing Surveys (CSUR)* 54, 10s (2022), 1–42.

[10] Alessandro D'Amelio, Vittorio Cuculo, and Sathya Bursic. 2019. Gender Recognition in the Wild with Small Sample Size-A Dictionary Learning Approach. In *International Symposium on Formal Methods*. Springer, 162–169.

[11] Giuliano Grossi, Raffaella Lanzarotti, and Jianyi Lin. 2016. Robust face recognition providing the identity and its reliability degree combining sparse representation and multiple features. *International Journal of Pattern Recognition and Artificial Intelligence* 30, 10 (2016), 1656007.

[12] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems* 30 (2017).

[13] Hu Han, Charles Otto, Xiaoming Liu, and Anil K Jain. 2014. Demographic estimation from face images: Human vs. machine performance. *IEEE transactions on pattern analysis and machine intelligence* 37, 6 (2014), 1148–1161.

[14] Anupama K Ingale, A Anny Leema, HyungSeok Kim, and J Divya Udayan. 2023. Automatic 3D Facial Landmark-Based Deformation Transfer on Facial Variants for Blendshape Generation. *Arabian Journal for Science and Engineering* 48, 8 (2023), 10109–10123.

[15] László A Jeni, Jeffrey F Cohn, and Takeo Kanade. 2015. Dense 3D face alignment from 2D videos in real-time. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, Vol. 1. IEEE, 1–8.

[16] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[17] Maksim Kuprashevich and Irina Tolstykh. 2023. MiVOLO: Multi-input Transformer for Age and Gender Estimation. *arXiv preprint arXiv:2307.04616* (2023).

[18] Sebastian Lapuschkin, Alexander Binder, Klaus-Robert Muller, and Wojciech Samek. 2017. Understanding and comparing deep neural networks for age and gender classification. In *Proceedings of the IEEE international conference on computer vision workshops*. 1629–1638.

[19] Gil Levi and Tal Hassner. 2015. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 34–42.

[20] Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE transactions on affective computing* 13, 3 (2020), 1195–1215.

[21] Shu Liang, Jia Wu, Seth M Weinberg, and Linda G Shapiro. 2013. Improved detection of landmarks on 3D human face data. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 6482–6485.

[22] Feng Lin, Yingxiao Wu, Yan Zhuang, Xi Long, and Wenyao Xu. 2016. Human gender classification: a review. *International Journal of Biometrics* 8, 3-4 (2016), 275–300.

[23] Xiaoguang Lu and Anil K Jain. 2006. Automatic feature extraction for multiview 3D face recognition. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. IEEE, 585–590.

[24] Gokhan Ozbulak, Yusuf Aytar, and Hazim Kemal Ekenel. 2016. How transferable are CNN-based features for age and gender classification?. In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 1–6.

[25] Sabrina Patania, Giuseppe Boccignone, Sathya Buršić, Alessandro D'Amelio, and Raffaella Lanzarotti. 2022. Deep graph neural network for video-based facial pain expression assessment. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. 585–591.

[26] Rasmus R Paulsen, Kristine Aavild Juhl, Thilde Marie Haspang, Thomas Hansen, Melanie Ganz, and Gudmundur Einarsson. 2018. Multi-view consensus CNN for 3D facial landmark placement. In *Asian Conference on Computer Vision*. Springer, 706–719.

[27] Pau Rodríguez, Guillem Cucurull, Josep M Gonfaus, F Xavier Roca, and Jordi Gonzalez. 2017. Age and gender recognition in the wild with deep attention. *Pattern Recognition* 72 (2017), 563–571.

[28] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. 2009. Fast point feature histograms (FPFH) for 3D registration. In *2009 IEEE international conference on robotics and automation*. IEEE, 3212–3217.

[29] Caifeng Shan. 2012. Learning local binary patterns for gender classification on real-world face images. *Pattern recognition letters* 33, 4 (2012), 431–437.

[30] Weijing Shi and Raj Rajkumar. 2020. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1711–1719.

[31] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1701–1708.

[32] Sergey Tulyakov and Nicu Sebe. 2015. Regressing a 3D face shape from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*. 3748–3755.

[33] Muhammad Umair and Muhammad Naqeeb Nazir. 2021. Classification of Demographic Attributes from Facial Image by using CNN. In *2021 International Conference on Artificial Intelligence (ICAI)*. IEEE, 68–73.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[35] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[36] Rajesh Verma, Navdha Bhardwaj, Pushap Deep Singh, Arnav Bhavsar, and Vishal Sharma. 2021. Estimation of sex through morphometric landmark indices in facial images with strength of evidence in logistic regression analysis. *Forensic Science International: Reports* 4 (2021), 100226.

[37] Mai Tuong Vi, Vinh Truong Hoang, Tram-Anh Nguyen-Thi, et al. 2021. Unsupervised gender prediction based on deep facial features. In *2021 Zooming Innovation in Consumer Technologies Conference (ZINC)*. IEEE, 1–4.

[38] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. 2019. Graph convolutional networks: a comprehensive review. *Computational Social Networks* 6, 1 (2019), 1–23.

[39] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. 2016. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3438–3446.

[40] Ziqing Zhuang, Douglas Landsittel, Stacey Benson, Raymond Roberge, and Ronald Shaffer. 2010. Facial anthropometric differences among gender, ethnicity, and age groups. *Annals of occupational hygiene* 54, 4 (2010), 391–402.