Perceptual Differences Between Men and Women: A 3D Facial Morphometric Perspective

Syed Zulqarnain Gilani and Ajmal Mian
School of Computer Science and Software Engineering
The University of Western Australia
Email: zulqarnain.gilani@uwa.edu.au

Abstract—Understanding the features employed by the human visual system in gender classification is considered a critical step towards improving machine based gender classification systems. We propose the use of 3D Euclidean and geodesic distances between biologically significant facial landmarks to classify gender. We perform five different experiments on the BU-3DFE face database to look for more representative features that can replicate our visual system. Based on our experiments we suggest that the human visual system looks at the ratio of 3D Euclidean and geodesic distance as these features can classify gender with an accuracy of 99.32%. The features selected by our proposed gender classification experiment are robust to ethnicity and moderate changes in expression. They also replicate the perceptual gender bias towards certain features and hence become good candidates for being a more representative feature set.

I. INTRODUCTION

The Human Visual System (HVS) is capable of interpreting a wealth of information from the face. One of the most important information transmitted by a face is the person’s gender. There are two challenging tasks in understanding the perceptual basis of how we classify facial gender: (1) to know the type of features employed by HVS in deciding the gender of a face and (2) to evaluate the possibility of improving gender classification algorithms based on these features. The facial gender is considered to be a continuum between the two extremes of masculinity and femininity. For this reason objective gender classification, using computer algorithms is a non-trivial task. Figure 1 shows 3D shaded images of the synthetic 3D face with gender varying from highly masculine to highly feminine.

Facial gender classification finds its applications in surveillance, human-computer interaction, demographic studies, content-based retrieval and targeted advertising [1]. It can also be used as a preprocessing step for person identification based on face biometrics. Current computer based facial gender classification algorithms are far behind the accuracy of human visual perception [2], perhaps, because there is a lack of understanding about the features employed by the HVS in performing this task.

Our literature survey shows that while there is a considerable effort in the psychology domain in understanding the perceptual nature of gender classification, significant effort in translating this understanding to improve computational algorithms is lacking. From a purely perceptual point of view, Bruce et al. [3] strived to find the particular information used by the human visual system in deciding the gender of a face. The authors performed five experiments on different sets of male and female; textured and 3D faces. They employed human observers to classify the gender of these faces in the first three experiments and to rate their masculinity in the last two. Faces were shown to observers in profile and oblique views with and without masks placed on the eyes, nose and chin regions of the face. The authors suggested that HVS takes into account the 2D, 3D facial structure, textural information and their interdependencies into account while deciding on gender. The authors also opined that the size and protuberance of nose contributes towards higher masculinity while the variation in the chin does not have a significant effect. In another perceptual study Yamaguchi et al. [4] analysed the contribution of various facial regions on gender classification. The authors recruited 72 male and 50 female participants and asked them to classify the coloured photographs of 26 male and 26 female subjects from Japanese decent. The authors suggested that it was possible to identify the gender solely on the basis of face and that the eyebrows and the face outline are important cues in identifying the gender.

From a computational point of view, gender classification accuracy yielded by algorithms has been used as the criteria for evaluating the significance of different features. A higher classification accuracy means better agreement on features between the classification algorithm and human visual system. In this connection, Burton et al. [2] looked at the information available to the HVS from a face to make fast decisions on its gender. In order to isolate the specific 2D and 3D features that are significant for gender classification the authors manually annotated 73 landmarks on a dataset of 91 male and 88 female faces yielding 2628 possible pairs of landmarks for feature extraction. However, acknowledging the computational complexity of processing such a huge amount of data, the authors used various subsets of these combinations in five different studies. Discriminant Function Analysis (DFA) for gender classification was performed using 2D and 3D Euclidean distances, ratios of these distances and angles between landmarks. The authors argued that perhaps the HVS takes into account a subset of 16 measurements to classify facial gender since these features result in a classification accuracy of 94%, which was the best amongst all studies.

Farkas et al. [5] identified a set of 23 biologically relevant anatomical facial landmarks to extract a set of Euclidean distances, ratios and angles for facial anthropometry. Since gender is a biological trait, it makes sense to focus on biologically significant landmarks for facial gender classification. A region based fully automated gender classification algorithm
was proposed by Hu et al. [6] using 15 automatically detected biologically significant facial landmarks which were used as features. These landmarks were then grouped into five gender discriminating regions of internal face, upper face, left eye, lower face and nose regions. Separate SVMs were trained on the features and the results were fused. They reported 94.3% accuracy, however, their database is private, consists of only 216 subjects of the same ethnicity and has no expression variations.

Researchers have also been interested in investigating the gender bias of expressions. It is suggested that certain expressions enhance the features of gender dominance in either the male or female gender [7], [8]. Hess et al. [8] asked 143 men and 156 women to classify the gender of a single avatar with varying expressions of anger, fear and happiness. The authors found that faces that displayed an angry emotion were more likely to be associated with males whereas faces that depicted the expressions of fear or happiness were associated with females. It was suggested that these expressions seem to selectively enhance those facial features that are pertinent to both the domain of emotions and gender. It would be intuitive to investigate whether this trend is replicated by using gender discriminating features found by computer algorithms. Such a scenario would strengthen the relationship between perceptual and computational gender classification models.

Based on our literature survey, we find that there is a significant gap in research in understanding the features employed by our visual system in classifying the gender. Relying on perceptual findings, psychologists have nominated certain features that discriminate between the two genders. However, computational techniques lack in optimally utilizing these features to achieve high gender classification accuracy. More specifically, no current technique simultaneously satisfies the following requirements: (1) full automation (2) high accuracy (3) use of biologically significant landmarks or features (4) validation on a large public dataset containing multiple ethnicities and facial expressions and finally (5) validation of perceptual bias of expressions towards a certain gender.

In this work, we propose a fully automatic facial gender classification algorithm in order to understand the perceptual basis of sexual dimorphism in human faces and the relationship of facial expressions to gender classification accuracy in men and women. We use minimum Redundancy maximum Relevance algorithm [9] combined with a forward selection wrapper targeting a linear classification algorithm to select a subset of Euclidean and Geodesic distances extracted between biologically significant landmarks. Separate experiments employing 3D Euclidean and geodesic distances on one of the largest public 3D face databases BU-3DFE [10] show that the human visual system looks at the ratios of these distances while classifying facial gender. With our selected features and using 10-fold cross validation, we achieve 99.32% gender classification accuracy. Further experiments on the multiple ethnicities of the database show that our algorithm is robust to variation in ethnicity. We investigate the gender bias of various expressions and show that our selected features replicate the perceptual trend reported by Hess et al. [8] and hence, may represent the human perceptual model for gender classification.

II. 3D FACE DATA AND LANDMARKS

We have experimented on 3D face data of the publicly available Binghamton University 3D Facial Expression (BU-3DFE) database [10]. The database comprises of 2500 scans of 100 subjects in frontal pose captured with the 3DMD face imaging system [11]. There are 25 scans of each subject in seven different expressions namely neutral, angry, disgust, fear, happy, sad and surprise. Except for the neutral expression each of the six expressions have four levels of intensity. BU-3DFE is a multiple ethnicity database having scans of subjects from six different races. Figure 3 shows all expression (less neutral) scans of one subject in the fourth level of intensity.

The dataset comes with manually annotated 83 feature points out of which we select 18 landmarks that are known to be biologically significant for clinical applications [5]. We have selected the facial landmarks that relate to the bony structure of the face which is effected by the ratio of testosterone to...


**Fig. 3.** Fourth level intensity expressions (anger, disgust, fear, happiness, sadness and surprise) of a white female subject in BU-3DFE database.

**Fig. 4.** Top: Process for automatically detecting Nasion (N) using curve analysis on BU-3DFE dataset. Bottom: Surface extracted within 30mm of feature point number 42. The peak on this surface corresponds to the nose tip. The detected landmarks are added to the existing set of 18 biologically significant landmarks.

The original landmark dataset is devoid of Nasion (N) and Pronasale (Prn). Nasion is the distinctly depressed point at the root of the nose that indicates the junction of the intranasal and the nasofrontal structures whereas Pronasale is the apex or the most protruded point on the nose also called the nose tip. We have automatically extracted these two landmarks using curvature analysis. To find the nasal bone, we extract the geodesic curve between the ground truth points number 37 and 48 and locate its peak. Next, we extract multiple geodesic curves on the nasal bridge within 10mm of the peak which correspond to the surface region at the junction of the intranasal and the nasofrontal structures. The negative peak on this surface corresponds to the Nasion. Automatic detection of the nose tip is straight forward. We extract the 3D surface within 30 millimetres of feature point number 42. The highest point on this surface is the nosetip (Prn). Fig. 4 details the procedure of locating these landmark.

Details of the database and landmarks are given in Table I.

### III. PROPOSED ALGORITHM

An overview of our gender classification algorithm is given in Figure 5 and its different components are explained below.

**A. Pre-processing**

BU-3DFE database comes with smooth hole filled models with 83 feature points defined on each model. We automatically detect the Nasion (N) and Pronasale (Prn) as already explained in section II. Next we correct the pose of each 3D face to a canonical form based on four landmarks (Ex(L), Ex(R), N and Prn). This step is required to eliminate any error due to pose in the extraction of geodesic distances.

**B. Feature Extraction**

We use two feature types, i.e. 3D Euclidean and the geodesic distances between the biologically relevant land marks. Given 20 landmarks on a 3D face, we extract twenty-five 3D Euclidean and 24 geodesic distances between them. Extraction of Euclidean distances is a straightforward task. We define geodesic distance as the length of the curve generated by orthogonal projection of the Euclidean line on the 3D facial surface. This is precisely the reason for normalising the pose of each 3D face as variation in pose can present a different surface to the viewing angle. Less curved distances like upper lip height (Sn-Sto) are modelled by a second order polynomial while more curved distances with multiple inflection points, like the biocular width (Ex-Ex) are modelled by higher order polynomials.

The motivation for using these feature types comes from the fact that they represent the sexual dimorphism of the face [15]. In clinical practice 3D Euclidean distances are used to measure a quantitative dimension for the morphological deviation from the normal face [16]. These distances have also been used to delineate syndromes [17]. Psychologists have used these features to measure the amount of facial masculinity/femininity [18]. On the other hand, studies suggest that geodesic distances may represent 3D models in a better way as compared to 3D Euclidean distances [19]. Gupta et al. [20] argue that algorithms based on geodesic distances are likely to be robust to changes in facial expressions. Figure 6 shows the 3D Euclidean and geodesic distance between left cheek (mouth corner) and exocanthion (eye corner) in neutral and extreme happy expressions of the same individual. The geodesic distance does not vary while there is a significant change in the Euclidean distance. Given these properties of 3D Euclidean and geodesic distances it seems appropriate to use them for facial gender classification of images with variations in expression.

**C. Feature Selection**

Feature selection is performed using the minimal redundancy maximal relevance (mRMR) algorithm packed in a forward-selection wrapper [9]. The algorithm first calculates the intrinsic information (relevance) within a feature and also the mutual information (redundancy) among the features to segregate different classes. Then it maximizes the relevance...
and minimizes the redundancy simultaneously. Let \( X \in \mathbb{R}^{m \times n} \) be the feature matrix with \( m \) observations and \( n \) features, \( F \) be the target reduced feature set and \( c \) be any arbitrary class from the set of classes \( C \), then relevance is defined by,
\[
D(F, c) = \frac{1}{|F|} \sum_{x_i \in F} I(x_i; c),
\]
and redundancy is defined by,
\[
R(F) = \frac{1}{|F|} \sum_{x_i, x_j \in F} I(x_i; x_j),
\]
where \( I(x; y) \) is the mutual information between \( x \) and \( y \). Maximal relevance and minimal redundancy is obtained by taking the maximum and minimum values of the above two equations respectively. The goal of simultaneously maximizing the relevance and minimizing the redundancy is achieved by maximizing the function \( \Gamma(D, R) \) where, \( \Gamma(D, R) = D - R \), or \( \Gamma(D, R) = \frac{D}{R} \). The former equation is the Mutual Information Difference (MID) and latter is the Mutual Information Quotient (MIQ) formulation of mRMR algorithm. Since our feature set is small we find the classification accuracy yielded by both formulations and use only the one giving the maximum accuracy on training data. The reduced number of candidate features \( k \) is selected by first obtaining \( n \) feature sets \( F_n \) using the mRMR sequential search (MID or MIQ formulation depending on which one gives better accuracy). More specifically \( F_1 \subset F_2 \subset \ldots \subset F_{n-1} \subset F_n \). Next we compare the classification accuracy for all feature subsets \( F_1, \ldots, F_k, \ldots, F_n \) \((1 < k < n)\) to find a range for \( k \) where the classification accuracy is maximum. Finally, we select a compact set of features \( p \) by exploiting the forward-selection wrapper [21]. Consequently, we obtain a feature set \( \{\Theta_1, \ldots, \Theta_p, \ldots, \Theta_k\} \) where \( 1 < p < k \) and we select the feature subset \( \{\Theta_1, \Theta_2, \ldots, \Theta_p\} \) which corresponds to the highest accuracy. Note that this is the most compact feature subset as \( 1 < p < k < n \).

**D. Gender Classification**

Next we train a Linear Discriminant Analysis (LDA) classifier using an exclusive set of training data. Let \( X_i \in \mathbb{R}^{p \times n_i} \) be the matrix of features of class \( i \) with \( n_i \) samples and \( p \) features as chosen in the feature selection step. LDA maximizes the ratio of \( \text{between-class scatter} \) to \( \text{within-class scatter} \). Between-class scatter is defined as, \( S_B = \sum_{i=1}^{c} n_i (\mu_i - \mu) (\mu_i - \mu)^\top \), and within-class scatter is defined as \( S_W = \sum_{i=1}^{c} \sum_{x_k \in X_i} (x_k - \mu_i) (x_k - \mu_i)^\top \), where \( \mu_i \) is the mean of all classes, \( \mu \) is the mean of class \( X_i \), and \( n_i \) is the number of samples in \( X_i \). Fisher [22] proposed to maximize the ratio between \( S_B \) and \( S_W \) relative to the projection direction by solving
\[
J(w) = \arg \max_w \frac{w^\top S_B w}{w^\top S_W w}.
\]

By differentiating the equation with respect to \( w \) and equating it to zero, we get \( S_W^{-1} S_B w = Jw = 0 \), which is a generalized eigenvalue problem and the eigenvector \( w^* \) of \( S_W^{-1} S_B \) is the desired optimal direction. Given the learnt LDA projection \( w^* \), a query face is classified into one of the two genders. The projection of feature vector \( x_q \) (of a face with unknown gender) on the LDA space is given by \( x_q^* = (w^*)^\top x_q \).

Gender classification is performed based on the distance between the \( x_q^* \) and the means of the projected classes \( \mu_1^* \) and \( \mu_2^* \) such that
\[
C_q = \begin{cases} 1 & \|\mu_1^* - x_q^*\|_2 < \|\mu_2^* - x_q^*\|_2, \\ 0 & \text{otherwise} \end{cases}
\]

IV. EXPERIMENTATION AND RESULTS

We perform a number of experiments to search for the best features that classify gender and investigate the robustness of these features in different scenarios. Experiments 1 to 3 are aimed at searching for the feature types that are possibly used by the human visual system in classifying the facial gender. Experiment-4 investigates the effect of ethnicity on the selected features. Finally, Experiment-5 investigates the gender bias of the six expressions in the database employing our selected features. The evaluation criterion for all experiments is the average gender classification accuracy with 10 fold cross validation. In each fold, all the scans of an identity appear in either the test set or training set and not in both. A detailed explanation of these experiments is given in the following subsections.

A. Experiment-1: 3D Euclidean distances

We classify gender using 25 3D Euclidean distances extracted between the 20 biologically significant landmarks as shown in Figure 2(b). Our proposed algorithm classifies 88.36% faces correctly as male or female (Table II). This result is far below the human accuracy threshold of 96% as reported by Burton et al. [2]. Hence, 3D Euclidean distances do not seem to be the feature of choice for the HVS in classifying gender. It is interesting to note that the forehead width (Ft-Ft), biocular width (Ex-Ex), nasal bridge length (N-Prn) and nasal tip protrusion (sbal-prn-sbal) are selected as the most differentiating features by our algorithm.
B. Experiment-2: Geodesic distances

In this experiment, we use 24 geodesic distances extracted between the 20 biologically significant landmarks to classify gender. These features classify facial gender with an accuracy of 90.92%, indicating their superiority over 3D Euclidean distances. However, being far below the human accuracy threshold, even geodesics alone do not seem to be the representative features for human visual system in gender classification. Once again forehead width (Ft-Ft), nasal bridge length (N-Prn) and nasal tip protrusion (shal-prn-shal) are selected as the most differentiating features by our algorithm. However, with geodesic distances, intercanthal width (En-En) is also selected as a gender discriminating feature.

C. Experiment-3: Combination of Euclidean and geodesic distances

In the last of the series of experiments to search for features that would help us understand the human visual system, we perform gender classification using the 24 geodesic and 25 3D Euclidean distances. Interestingly, the classification accuracy with these features comes out to be 99.32%, shooting well above the human accuracy threshold of 96% [2]. Analysis of the most discriminating features shows that our algorithm selects both 3D Euclidean and geodesic distances of the forehead width, nasal bridge length (N-Prn) and nasal tip protrusion, intercanthal width and mandibular width (jaw width) in achieving this high accuracy. Figure7 shows the selected features on the 3D shaded faces of some of the subjects.

D. Experiment-4: Effects of ethnicity

Using the discriminating features found in Experiment-3 we proceed to investigate their robustness on ethnicity variation. The BU-3DFE database consists of subjects from six ethnicities. Gender and ethnicity distribution is given in TableIII. Each time we train our classifier on subjects of one ethnicity (all expressions) and test on the rest of the ethnicities. We exclude Black, Indian and Middle-East Asian ethnicities from the training set as they are heavily biased towards one gender. However, we do not exclude them from test sets.

This makes our experimental setup ever more challenging. Results depicted in Table IV show that our selected features are invariant to ethnicity.

E. Experiment-5: Gender bias of the expressions

In the final experiment we train the gender classification algorithm on the neutral expression scans of the subjects in BU-3DFE using the features found in Experiment-3. Next we perform gender classification on each expression (all intensity levels) separately and note the misclassification rate for males and females. Our results show that male misclassification rate increases in the expressions of disgust, fear, happiness and surprise whereas female misclassification increases in anger only. The expression of sadness does not have any significant effect on the relative misclassification rate of either gender. Figure 8 shows the misclassification rate for each gender in the presence of varying intensities of the six expressions.

V. Analysis

In the first three experiments we have searched for the appropriate features that might be used by the human visual system for classifying the facial gender. With the 3D Euclidean distances we see that the gender classification accuracy of 88.36% is significantly below the human perceptual accuracy as well as the classification accuracy (94%) of Burton et al. [2]. The obvious reasons seems to be the choice of features. While we have selected only the 3D Euclidean distances, Burton et al. chose a combination of 2D/3D Euclidean distances and ratios and angles between these distances. Experiment-2 shows that geodesic distances between biologically significant

![Fig. 7. 3D Euclidean (top row) and geodesic (bottom row) features selected by our algorithm to classify facial gender shown on neutral, extreme disgust and sad expression faces.](image)

![Fig. 8. Male and Female misclassification rates (in %age) for four intensity levels of six expressions. The expression of anger supports males whereas disgust, fear, happiness and surprise favour females.](image)
landmarks yield a better gender classification accuracy, though still significantly below the human accuracy threshold. Hence, only 3D Euclidean or geodesic distances do not seem to represent the features selected by human visual system for classifying facial gender. In Experiment-3, where a combination of 3D Euclidean and geodesic distances are used as features, the classification accuracy becomes par excellence. Our proposed features classify gender with an accuracy of 99.32%. A closer look at the selected features (Figure7) suggests that it is the ratios of certain critical Euclidean and geodesic distances that yield a high classification accuracy, and perhaps these represent the human visual model in gender classification.

Equipped with strong gender differentiating features we investigate the effects of ethnicity on gender classification. While most of the previous studies [2]–[4], [6], [23], [24] have been performed on a single ethnicity our findings suggest that our selected features are robust to different ethnicities.

Finally in Experiment-5, we investigate the gender bias of all six expressions available in BU-3DFE dataset. Our results corroborate the perceptual findings [7], [8] and suggest that angry looking females in 3D images are more likely to be misclassified as males. In all the four intensities of anger, none of the males is misclassified as a female whereas, females are increasingly misclassified as males as the expression intensifies. Similarly, males expressing fear, happiness and less investigated emotions of disgust and surprise are increasingly misclassified as females as the expression intensities. The expression of sadness exhibits a neutral bias where the relative misclassified as females as the expressions intensify. The investigated emotions of disgust and surprise are increasingly of the males is misclassified as a female whereas, females misclassified as males. In all the four intensities of anger, none angry looking females in 3D images are more likely to be classified gender can replicate the perceptual gender bias trend of expressions and can hence represent the human visual system in classifying gender.

VI. CONCLUSION

In this paper, we looked for facial features that represent the human visual system in gender classification. We performed five experiments for this purpose. As a result of the first three experiments we showed that the ratio of 3D Euclidean and geodesic distances extracted between biologically significant facial anatomical landmarks can replicate the perceptual system as these features classify gender with an accuracy of more than 99%. Our results from the second experiment suggest that the proposed features are robust to ethnicity. Experiment-4 provides the basis for our findings that the selected features that classify gender can replicate the perceptual gender bias trend shown by certain expressions and the features are also robust to moderate changes in expression. Based on our experiments we suggest that, the ratio of 3D Euclidean and geodesic distances can optimally represent the human visual system in classifying gender.

ACKNOWLEDGEMENT

Syed Zulqarnain Gilani is funded by the International Postgraduate Research Scholarship (IPRS). This research was also supported by ARC grant DP110102399 and the UWA FECM grant.

REFERENCES