Sex Classification of Face Images using Embedded Prototype Subspace Classifiers

Anders Hast[1]
[1] Department of Information Technology
Uppsala University
Centre for Image Analysis
SE-751 05 Uppsala, Sweden
anders.hast@it.uu.se

ABSTRACT

In recent academic literature Sex and Gender have both become synonyms, even though distinct definitions do exist. This give rise to the question, which of those two are actually face image classifiers identifying? It will be argued and explained why CNN based classifiers will generally identify gender, while feeding face recognition feature vectors into a neural network, will tend to verify sex rather than gender. It is shown for the first time how state of the art Sex Classification can be performed using Embedded Prototype Subspace Classifiers (EPSC) and also how the projection depth can be learned efficiently. The automatic Gender classification, which is produced by the *InsightFace* project, is used as a baseline and compared to the results given by the EPSC, which takes the feature vectors produced by *InsightFace* as input. It turns out that the depth of projection needed is much larger for these face feature vectors than for an example classifying on MNIST or similar. Therefore, one important contribution is a simple method to determine the optimal depth for any kind of data. Furthermore, it is shown how the weights in the final layer can be set in order to make the choice of depth stable and independent of the kind of learning data. The resulting EPSC is extremely light weight and yet very accurate, reaching over 98% accuracy for several datasets.

Keywords

Sex and Gender Classification, Subspaces, Embedded Prototype Subspace Classification, Face Recognition.

1 INTRODUCTION

Sex or Gender classification is one important task in the field of face recognition (FR) and it will be shown how this can be done efficiently using *Embedded Prototype Subspace Classification* (EPSC) [Has22, HV21, HLV19, HL20, HV21], which has already been proven to be able to classify datasets of various kinds, such as digits, words and objects.

Another important contribution to EPSC in this paper, is to show how the projection depth can be learned and how the weights in the final layer can be set in order to make sure that the algorithm is stable and that the results are always reliable, regardless of how the depth parameter is set for unknown data to be classified.

1.1 Sex or Gender?

Sex and Gender has increasingly become synonyms and the recent research papers about how to use FR to determine whether there is a male or a female in the image, generally use the word *Gender*, just as many other academic papers tend to do in their titles [Hai04] nowadays. However, there is a distinct difference between the words *Sex* and *Gender*, and the Council of Europe gives several definitions on one of their websites [oE23]. To make a long story short, *Sex* generally refers to "biological differences between males and females", while *Gender* is presented as a "social, psychological and cultural construct".

In this paper the question about sex and gender will be handled in a very simple way, by just using the labels "Male" and "Female" provided in the datasets. Hence, the classification has already been done in some way, and then the task is for any computer algorithm to determine the class based on cues such as face geometry, facial hair etc, as explained in the following sections. The main question is which cues to use and whether humans use other cues than Machine Learning (ML) algorithms for sex and gender classification. It will be argued that when using face feature vectors (FFV) from

FR, they are predominantly based on the geometry of the face, regardless of facial expression and outer attributes such as facial hair, makeup or hair-cuts, which otherwise could reveal the gender of the person at hand.

1.2 Cues

Hoss et al. [HRGL05] found that high masculinity in male faces, but not high femininity in female faces, facilitated sex classification when showing face images to both adults and children. And they also found that, independently of masculinity/femininity, attractiveness affected not only the accuracy of the sex classification, but also the speed.

Interestingly, Bruce et al. [BBH*93] found that cues from features such as eyebrows, and skin texture, play a more important role when humans are deciding whether faces are male or female, than cues from such things as hairstyle, makeup, and facial hair. Moreover, O'Toole et al. [OPD96] found that female observers were more accurate at classification of sex than were male observers, on both Caucasian and oriental faces.

Obviously, humans use different cues to determine sex and gender, both outer attributes but also the geometry itself, which is related to masculinity and femininity.

1.3 Geometry or Attributes?

The main question then is, which of those two concepts will FR algorithms pick up on? A Convolutional Neural Network (CNN) will probably pick up on both facial attributes, such as hairstyle, facial hair but also on geometrical aspects such as size of chin and nose, etc. The reason is simply, because they are all visual cues present to different degrees in face images.

It is quite obvious that *Sex* could have an impact on all these cues, while *Gender*, i.e. a persons own perception about gender, cannot change the geometrical aspects, since they are predominantly the results of a persons sex. However, depending on a persons *Gender* they have indeed power to change the other attributes, such as facial hair, makeup and haircut, which the surrounding world would perceive as typically female or male (or neither).

FFVs on the other hand are constructed so that the FR software can be used to recognise a person, regardless of the aforementioned outer attributes, which might reveal the persons gender. Hence, the FFV are more stable in that sense, and therefore obviously tend to pick up on the geometry, which is more related to *Sex*. So for this reason, the title contains the word *Sex Classification*, rather than *Gender Classification*. Nonetheless, as stated before, most researchers will understand both as determining whether the person present in the image is a male or a female.

2 RELATED WORK

Burton et al. [BBD93] compared the human ability to determine sex of persons showing the face only, wearing swimming caps to conceal their hair (outer attribute), compared to a computer approach based on the face geometry, and found that humans could reach an accuracy of 96% and that the computer was at the time approaching human performance of 94% accuracy. These experiments from 30 years ago underlines the importance of face geometry for determining sex.

Golumb et al. [GLS90] devised "SexNet", which was a neural network, working on aligned and scaled face images, and had an average error rate of 8.1% compared to humans, who averaged 11.6% on that particular dataset. Hence, this network would pick up on both the geometry and outer facial cues like facial hair.

Mäkinen and Raisamo[MR08] gives an overview of different gender classification methods with a varying accuracy between 76.87% and 86.54%, on the FERET [PWHR98] and IMM [NLSS04] datasets.

Gong et al. [GLJ20] proposed a group adaptive classifier for face recognition, which is designed to customize the classifier for each demographic group, and automatically learns where to use adaptive kernels in a multilayer CNN. They obtained an accuracy for FR of 98.19% using a 5-fold cross-validation on 8 groups of the Racial Faces in the Wild (RFW) dataset [WDH*19]. However, the accuracy for gender classification was only 85%.

Gil and Hassner [LH15] achieve an accuracy of 86.8% using deep CNN's on the Adience Benchmark [EEH14], on which others have reached up to 91% [SBLM17].

Acien et al. [AMVR*19] used the Labeled Faced in the Wild (LFW) dataset [HRBLM07], which will be used also in this paper. They achieved a 94.8% accuracy using VGGFace and 89.01% using ResNet50, where a separate layer was added in both networks to classify gender. It can be noted that the dataset is ethnically mixed between Caucasians, Asians and Blacks.

Sumi et al. [SHIA21] gives an overview of several other works for gender classifcation, where no algorithm reaches over 97% accuracy. They report themselves an average training and test accuracy of 90% and 83.5%, respectively on the Nottingham Scan Database, which will also be used in this paper.

Both Liu et al. [LLWT15] and Ranjan et al. [RPC16] report results from several implementations trying to classify gender on the CelebA dataset, which is another dataset that will be used in this paper. The accuracy spans from 90% up to 98% depending on which method is being used.

2.1 3D Faces

Interestingly, Abbas et al. [AHM*18] proposed a set of facial morphological descriptors, based on 3D geodesic path curvatures between two key landmarks in 3D faces, obtained from 3D scanning. Hence, it would only pick up on geometry and not on outer facial cues, like facial hair. They achieved a gender classification accuracy of 88.6% using a combination of geodesic distances between landmarks and the new geodesic path features.

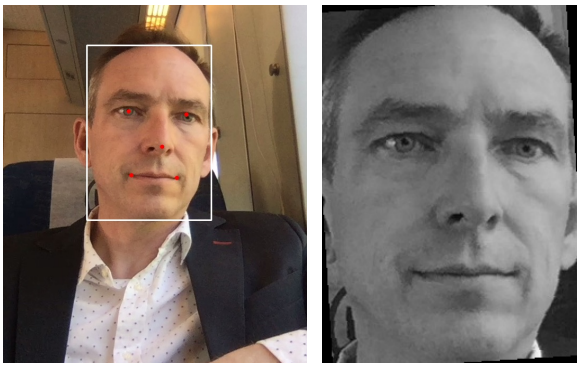
Gilani et al. [GRS*14] investigated how the the human cognitive system uses geometric features in perceiving the degree of masculinity/femininity in 3D faces. Their results suggested that humans use a combination of both Euclidean and geodesic distances between biologically significant landmarks of the face for determining gender. Gilani and Milan [GM14] used this approach on 3D scanned faces and obtained an accuracy of 99.32

It should be noted though that the proposed method can not be easily compared to 3D methods, since only 2D face images are used from FR software. However, the important lesson here is that geometry was concluded in both mentioned papers as an important cue for differing between masculine and feminine faces.

3 BACKGROUND

The main advantage of EPSC compared to many deep learning based methods [Sha18] is that EPSC is shallow to its nature, with no hidden layers, and therefore it does not require powerful GPU resources in the training process. Recently, the main idea of backpropagation used by most neural network based approaches, was questioned by Geoffrey Hinton [Hin22] who proposed another alternative in the *The Forward-Forward Algorithm*. Interestingly, the EPSC does not use backpropagation at all and learns only from the feature vectors using PCA, and from the embedding of feature vectors, using dimensionality reduction techniques like t-SNE, UMAP or SOM [HV21]. Consequently, subspaces are created from each cluster [KLR*77, OK83], which constitutes a set of neurons that are specialised on identifying the class variation captured in that cluster. However, in this paper, only one set of neurons, i.e. subspace, will be used for each class, as it turns out to be more efficient for the purpose of sex/gender classification.

Obviously, EPSC does not always outperform the state-of-the-art deep learning approaches when it comes to accuracy. However, Both learning and inference will generally be much faster, due to its simplicity and compactness. Moreover, both the learning and classification processes are inherently easy to both interpret [Kri19, CPC19] and explain [ADRS*19, GSC*19, CPC19], as well as they are easy to visualise. Furthermore, Deep learning does not always solve a problem better than classical machine learning algorithms [DDD*23]. Hence, there are several reasons to look at fast, and sustainable computing alternatives.



(a) Detected face (white box) with landmarks (red dots) (b) Face after alignment

Figure 1: Illustration of automatic face detection and alignment using landmarks.

4 FACE FEATURE VECTOR EXTRACTION

In this paper the *InsightFace* [Ins23] pipeline is used, which is an integrated Python library for 2D and 3D face analysis. *InsightFace* efficiently implements a rich variety of state of the art algorithms for both face detection, face alignment and face recognition, such as *RetinaFace*. [DGZ*19]. It allows for automatic extraction of highly discriminative FFVs for each face, based on the Additive Angular Margin Loss (ArcFace) approach [DGXZ19].

Face detection algorithms have come a long way since the simple, but efficient Viola-Jones detector [VJ01]. As an example, Figure 1a shows *RetinaFace* can perform automatic face detection that find the white bounding box of the face. Red landmarks are computed that can be used to align the face as shown in 1b.

The FFV, also known as embedding, will have length 512 when using *InsightFace* and the provided *buffalo_l* model, which is the default model pack in latest version of *InsightFace*. Other approaches such as *DeepFace* [TYRW14], *CosFace* [WWZ*18] *FaceNet* [SKP15], *SphereFace* [LWY*17], or [WZLQ16, SJ19], just to mention a few, could also have been used to produce FFVs for the proposed approach, as well as some of the ones mentioned in section 2.

5 DATASETS

In the automatic FR process, images where faces were not properly recognised or images with more than one face where removed. The remaining *face images*, i.e. an image containing one face and producing an FFV

to be processed further, were kept as shown in table 1. This process is referred to as the automatic selection in the following description of the datasets. For some datasets, it was required that several face images of the same person, covering several ages (decades) where chosen, while for others no such selection was done. In this way, quite different datasets are at hand and can be evaluated. Since the datasets used often are subsets of the original dataset, they will be referred to as explained below.

5.1 AgeDB

The *AgeDB* dataset [MPS*17] contains 16,516 images. Of those, 9826 face images were extracted so that each person depicted had about 36 face images on average covering at least three different age decades. Moreover, it was required that each person included should have at least 30 face images. This will ensure that there are several face images of the same person at different ages, or decades, and not only for one. Hence, it will make it possible to verify that the sex classification works for different ages.

5.2 CASIA

Since the *CASIA-WebFace* [YLLL14] is rather large as it contains around 500*k* images, a much smaller subset was extracted containing 65579 face images, which is still quite large compared to some of the other datasets used. Nevertheless, a similar approach was used as for AgeDB, resulting in more than 50 face images per person on average.

5.3 LFW

The *Labeled Faces in the Wild* (LFW) dataset [HRBLM07, LM14] contains 13,233 images, with 5749 different individuals. It is unbalanced as 1680 people have two or more images and the remaining 4069 have just a single image in the database. After FR, 10792 face images where selected automatically. No further selection was done, i.e. no requirements of age groups.

5.4 CelebA

The *CelebA* dataset [LLWT15] contains 202,599 images of 10,177 persons. The automatic FR extraction resulted in 200,096 face images where kept. No further selection was done.

5.5 UTKFace

The *UTKFace* datset [ZYQ17] contains 23,709 images, where 23,685 face images where kept after the automatic FR extraction. This set contains a wast age range from 0 to 116 years old, making it rather challenging for sex classification. Moreover, there are quite many images with watermarks, that could influence the face recognition. Anyhow, no further selection was done.

Table 1: Databases used, with the total number of face images, number of women and Men.

Database	Total	Women	Men
AgeDB	9826	4071	5755
CASIA	65579	24077	41502
LFW	10792	2410	8382
CelebA	200096	116985	83111
UTKFace	23685	11308	12377
NSD	100	50	50

5.6 NSD

The *Nottingham Scans Dataset* (NSD) was included as it is rather different from the other dataset, but also because it has been evaluated before [SHIA21]. It only contains 100 face images, all of different persons. However, it is totally gender balanced, with 50 face images of each sex. As can be seen from the table, the other datasets vary quite a lot when it comes to this aspect, which will affect the accuracy, both when it comes to training as well as classification.

6 SUBSPACE CLASSIFICATION

Subspace Classification in pattern recognition was introduced by Watanabe et al. [WP73] in 1967 and was later further developed by Kohonen and others [WLK*67, KLR*77, KO76, KRMV76, OK88]. The following mathematical derivation follows from Oja and Kohonen [OK88] and Laaksonen [Laa07].

Every face image to be classified is represented by a FFV $\mathbf{x}$ with m real-valued elements $\mathbf{x}_j = \{x_1, z_2 ... x_m\}, \in \mathbb{R}$, such that the operations take place in a m-dimensional vector space $\mathbb{R}^m$. In this paper m is equal to the FFV length, i.e. 512. Any set of n linearly independent basis vectors $\{\mathbf{u}_1, \mathbf{u}_2, ... \mathbf{u}_n\}$, where $\mathbf{u}_i = \{w_{1,j}, w_{2,j} ... w_{m,j}\}, w_{i,j} \in \mathbb{R}$, which can be combined into an $m \times n$ matrix $\mathbf{U} \in \mathbb{R}^{m \times n}$, span a subspace $\mathscr{L}_\mathbf{U}$

$$\mathscr{L}_\mathbf{U} = \{\hat{\mathbf{x}} | \hat{\mathbf{x}} = \sum_{i=1}^{n} \rho_i \mathbf{u}_i, \rho_i \in \mathbb{R}\} \tag{1}$$

where,

$$\rho_i = \mathbf{x}^T \mathbf{u}_i = \sum_{j=1}^{m} x_j w_{i,j} \tag{2}$$

Classification of a feature vector can be performed by projecting $\mathbf{x}$ onto each subspace $\mathscr{L}_{\mathbf{U}_k}$. The vector $\hat{\mathbf{x}}$ will in this way be a reconstruction of $\mathbf{x}$, using n vectors in the subspace through

$$\hat{\mathbf{x}} = \sum_{i=1}^{n} (\mathbf{x}^T \mathbf{u}_i) \mathbf{u}_i \tag{3}$$

$$= \sum_{i=1}^{n} \rho_i \mathbf{u}_i \tag{4}$$

$$= \mathbf{U}^T \mathbf{U} x^T \tag{5}$$

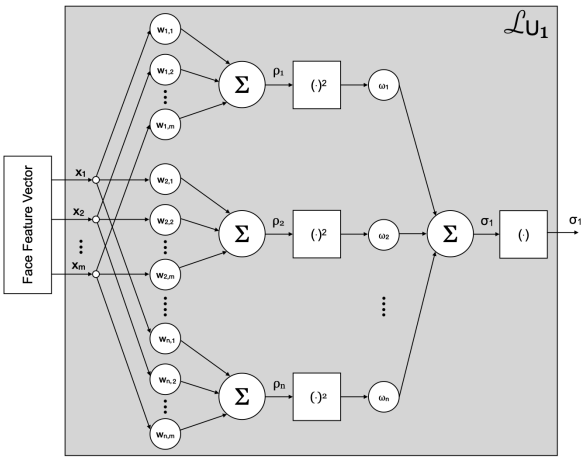


Figure 2: Illustration of the neural net that constitutes one of two subspaces forming the EPSC for Sex Classification. Neurons compute the response ρ from weights w, using the quadratic response function $(\cdot)^2$. In the second layer, the weights ω and the linear response function $(\cdot)$ are used.

By normalising all the vectors in $\mathbf{U}$, the norm of the projected vector can be simplified as

$$||\hat{\mathbf{x}}||^2 = (\mathbf{U}x^T) \cdot (\mathbf{U}x^T) \tag{6}$$

$$= (\mathbf{U}x^T)^2 \tag{7}$$

$$= \sum_{i=1}^{n} {\rho_i}^2 \tag{8}$$

Therefore, the feature vector $\mathbf{x}$, which is most similar to the feature vectors that were used to construct the subspace in question $\mathscr{L}_{\mathbf{U}_k}$, will thereby have the largest norm $||\hat{\mathbf{x}}||^2$.

The parameter n discussed above is the projection depth that will be learned, which is discussed in section 7.

In order to construct subspaces, a group of prototypes need to be chosen for each subspace, which is done by the embedding obtained from some dimensionality reduction method such as t-SNE [MH08], UMAP [MH18] or SOM [Koh82]. However, for classification of sex, it was found that one subspace per class was sufficient, and hence all feature vectors for males are used to construct one subspace, and all feature vectors for females are used for the second subspace.

In general, subspace classification can be regarded as a two layer neural network [HLV19, OK88, Laa07], where the weights are not learned using time consuming backpropagation. Instead, all weights are mathematically defined through Principal Component Analysis (PCA) [Laa07]. The resulting Neural Network for one subspace in the EPSC is shown in Fig. 2. The output of the two subspaces are then compared via the *argmax* function do determine which class, male or female, is at hand.

The neurons compute the response ρ using the quadratic response function $(\cdot)^2$, commonly referred to as an activation function. The mathematics of subspaces defines it to be quadratic, since it is deduced from computing the norm as in equation (8). The weights ω in the final layer should, according to the definition of the dot product, all be set to 1. However, it will be shown how it can be changed to make the classification more stable. For the same reason, the response function in the final layer is by definition linear, which still makes sense because it is the output to *argmax*.

7 LEARNING THE PROJECTION DEPTH

As stated earlier, the variable n in equation (1) is the projection depth. It tells how many dimensions in the subspace that are actually used when computing the projected vector $\hat{\mathbf{x}}$ in equation (3).

The projection depth needs to be determined for each type of classification task, and will somehow depend on number of classes and the data at hand. As an example, it was reported to be 28 for MNIST [HLV19] and 6 for the Esposalles word dataset [Has22]. Experimentally it was noted that rather large values for n was necessary for sex classification, and it depended quite a lot on which of the aforementioned datasets were used for training. The obvious way to learn the depth, is to vary it from 1 to m and find the optimal accuracy. Here $m = 512$ for the face feature vectors obtained from *InsightFace*.

Since the projection into the subspace can be regarded as a reconstruction of the input vector using the vectors in the subspace, it can be generally be observed that using a few vectors, i.e. small depth, gives more errors in the subspace reconstruction. On the other hand, using too many vectors, i.e. large depth, help in generating a *near perfect reconstruction*, making it impossible to suggest which subspace gives the strongest response.

Therefore, it is reasonable to deduce that the initial vectors in the PCA are more important than the subsequent ones, and therefore different weights ω could be applied. Referring to Fig. 2, let σ be the response output function from every subspace projection $\mathscr{L}_{\mathbf{U}}$, then a closed-form solution of each σ is computed as a weighted sum, presented in the following equation:

$$\omega_i = 1 - \left(\frac{i-1}{n}\right)^2 \tag{9}$$

where i varies from 1 to n. This decaying or dampening function generated the best results in general. The quadratic decay makes sure that the initial dimensions will use an ω closer to one, while the very last dimensions will use an ω closer to zero. Hence, the negative

effect of *near perfect reconstruction* is avoided as the trailing dimensions are hardly used at all.

The main reason for doing this is not only that the classification accuracy actually increases, but rather to avoid using a preset projection depth that accidentally would cause *near perfect reconstruction* for some datasets. Hence, the best projection depth for different kinds of datasets could be chosen with the reassurance that it will always be reliable and never cause the accuracy to drop substantially. All these claims will be proven in the next subsection where the experiments are explained.

7.1 Experiments

First three sets where chosen for learning, AgeDB, CASIA and LFW. Since the EPSC does not perform backpropagation, no epochs are performed. Instead the learning is simply performed by dividing the set for training into one cluster for all males and another one for all females. Only one subspace for each class is then created using PCA, since it was experimentally noted that not much was gained by dividing these subspaces further using, for an example t-SNE as proposed earlier [HLV19]. Then the projected depth is learned by finding the optimal depth using the validation set.

Two different approaches were deployed for computing the best projection depth. The first approach divided each set into a learning set (60%) and a validation set (40%). Here the datasets where split on person, so that the same person where only present in one of the splits. Furthermore, it splits on sex so that there are both 60% of the males and females in the set for learning and 40% of each sex for validation,

The second approach was simply to learn from one of the three sets and validate on each of the other sets. Hence, using the three aforementioned datasets as training sets, and validations sets, one at a time, six different permutations are possible.

In each experiment, the optimal depth for sex classification using the FFVs as input to the EPSC, was determined. The proposed dampening function in equation (9) to set the weights ω was used and then 100 runs where performed using bootstrapping on three datasets at a time, as shown in table 2, where 60% was used as sets for training and the remaining 40% as validation sets. The result can be compared to table 3, where the average accuracy for the validation set, when computing the sex with the provided model from *InsigtFace*. The tables shows the accuracy for the two classes, and the Macro Average Arithmetic, which is just the mean of the two classes. This is done to avoid the effect of the the fact that several of the datasets are heavily imbalanced and that could have a large impact on methods that are better to find one sex than the other. The MAA

Table 2: Average accuracies for 100 runs using bootstrapping with a 60/40 split and depth=360.

Database	Women	Men	MAA
CASIA	0.9995	0.9976	0.9986
AgeDB	0.9995	0.9907	0.9951
LFW	0.9900	0.9913	0.9907

Table 3: Accuracy for different datasets when the gender is classified by the face recognition model.

Database	Women	Men	MAA
CASIA	0.6824	0.7037	0.6930
AgeDB	0.8018	0.7293	0.7655
LFW	0.5651	0.6924	0.6288

Table 4: Optimal Depth for different datasets running 100 times using 60/40 split and bootstrapping. The mean optimal projection depth is 359.

Split	CASIA	AgeDB	LFW	Mean
60/40	361	373	342	359

is generally defined as the arithmetic average of the partial accuracies of each class:

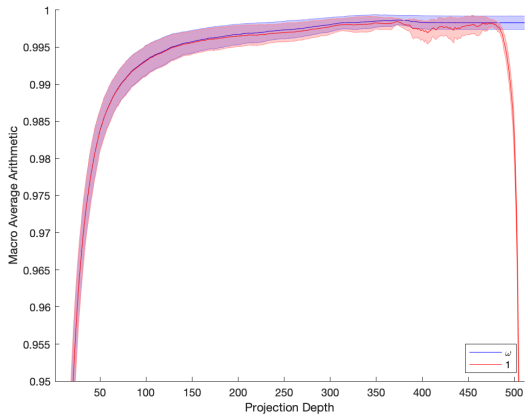
$$MAA = \frac{\sum_{i=1}^{N} ACC_i}{N} \tag{10}$$

where $N = 2$ for gender classification, as it the datasets have only two genders labeled. Hereby we avoid any discussions about what is actually gender, and how it relates to biological sex and perceived gender.

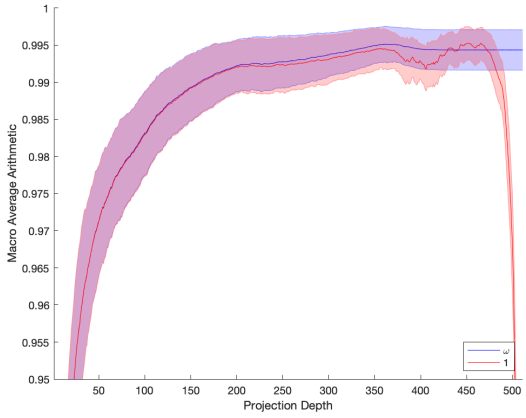
In the first approach it was chosen to compute the the MAA for sex classification using the Bootstrapping method [Koh95, KW96], with stratification because the data is imbalanced, which means that the bootstrap sample is taken from the whole set by using *sampling with replacement*. The experiments where conducted 100 times for each data split, varying the permutations randomly.

The results of the experiments are shown in Fig. 3 and are shown as a so called shaded bars graph, where the MAA is shown as a red curve with its shaded error (standard deviation) for $\omega = 1$. It can noted that for high projection depths *near perfect reconstruction* is reached and the MAA drops rapidly. While for the blue curve, using equation (9) for computing ω, the curve flattens out. The latter ensures that the projection depth can be set to high values, without the dangers of reaching *near perfect reconstruction*.

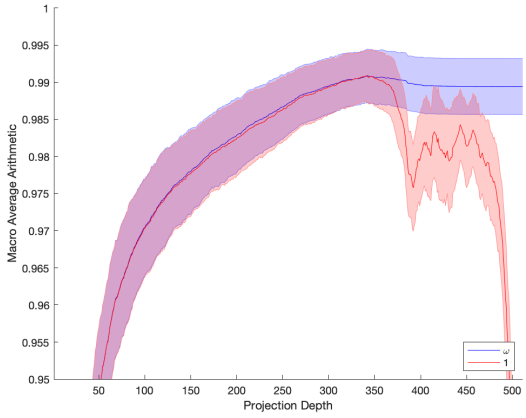
The optimal projection depth for the different datasets are reported in table 4. The optimal depth was computed when using ω in equation (9). The mean value is 359, which would still give a close to optimal MAA for all three datasets.



(a) CASIA



(b) AgeDB



(c) LFW

Figure 3: Macro Average Arithmetic for different depth of projection with shaded error (standard deviation), comparing dampening function ω (blue) for weights and using 1 as weight (red). Note how the former helps lifting the curve for larger projection depths in all three datasets

Table 5: Optimal Depth for different datasets for training and validation. The mean of all permutations, except those on the diagonal, is 360.

Dataset	AgeDB	CASIA	LFW
AgeDB	249	393	382
CASIA	328	395	352
LFW	375	328	347

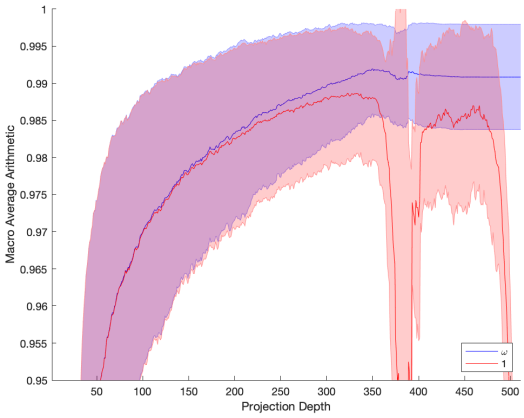


Figure 4: Macro Average Arithmetic for the mean of the 6 permutations of datasets with different depth of projection with shaded error (standard deviation), comparing dampening function ω (blue) for weights and using 1 as weight (red). Note how the former helps lifting the curve for larger projection depths on average for the datasets.

In Table 5 the optimal depth for all 6 possible permutations of training and validation sets is reported. Since using the same set for training and validation yields much better results, the results on the diagonal were not used when computing the mean equal to 359.6667. Once again ω was set as in equation (9).

Fig. 4 tells the mean of all six experiments, and clearly shows that the use of the dampening function for ω lifts the MAA curve and gives better accuracy than for $\omega = 1$.

The conclusion from both types of experiments is that a projection depth can safely be set around 360 for sex classification using different datasets for learning.

8 RESULTS

After concluding that 360 turns out to be a good value for the projection depth in the experiments, using different validation sets, the MAA was computed using this projection depth on different datasets for leaning and testing. Table 6 compares the MAA for using the EPSC approach and what accuracy is obtained by using the gender classificaion model provided by *InsightFace*. It can be noted that EPSC produce state of the art

Table 6: MAA for Classificaion of Sex, using one dataset for learning and another one for testing.

Database	AgeDB	CASIA	LFW	CelebA	UTKFace	NSD
AgeDB	*0.9994*	0.9921	0.9822	0.9596	0.8847	0.9900
CASIA	0.9960	*0.9994*	0.9860	0.9769	0.9007	0.9900
LFW	0.9963	0.9977	*0.9976*	0.9737	0.9065	**1.0000**
CelebA	0.9966	**0.9989**	0.9885	*0.9893*	**0.9152**	0.9500
UTKFace	**0.9974**	0.9979	**0.9897**	**0.9813**	*0.9488*	**1.0000**
InsightFace	0.7593	0.6959	0.6640	0.6887	0.5700	0.5807

results for sex classification and also outperforms the deep learning model.

9 DISCUSSION

The question about a persons *Sex* is easier to understand than *Gender* since the former is something that usually is assigned at birth, even if it sometimes can be a difficult task because of different disorders. The latter on the other hand can be chosen later in life. Nonetheless, as stated before, these two words have become synonyms and are also used as such such in this paper. However, it is closer at hand to talk about *Sex* classification for the method presented in this paper. The reason is that since FFVs are used, which are more closely related to the geometric features in a face, and therefore theoretically should be invariant to facial hair, hair style, makeup etc. Hence FFVs captures the geometry, which depends on genetics, rather than facial hair, hair-style, makeup etc, which all can be chosen. In any case, no political stance is taken in this paper about this matter. Furthermore, most datasets do not reveal on what grounds the sex or gender was chosen, so there is no other choice than trusting the labels and classify on them. However, the proposed algorithm, will lean towards classifying *Sex* rather than *Gender*, while many CNN based algorithms might pick up on both geometric features as well as outer features such as facial hair, hair-style and makeup, since these are the things that are visible in the face images.

The EPSC achieves state of the art classification on hard datasets. The way the dampening function is formulated in equation (9), which is used to set the weights ω is one important contribution in this paper. It makes the projection depth variable n less sensitive, and makes the EPSC reliable compared to not using the dampening function.

Looking at at the results in table 6 one can note that the proposed approach, using FFVs as input to EPSC performs close to or precede the state of the art. Acien et al. [AMVR*19] [HRBLM07] achieved a 94.8% accuracy using VGGFace on the LFW dataset, while the EPSC achives well over 98%.

For the CelebA dataset, both Liu et al [LLWT15] and Ranjan et al. [RPC16] reported the results from several implementations with an accuracy spanning from 90% up to 98%. The table shows that the EPSC is close to 98% or over depending on what dataset was used for learning.

It should be noted though that some methods used a split from the same dataset for training, validation and testing. Hence, the results cannot be compared straight on, but rather gives an indication on how well the proposed algorithm works compared to the state of the art algorithms. Furthermore, the learned projection depth could be set differently depending on the data at hand, but here it was chosen to learn it from three datasets. Learning from the same dataset when doing some kind of cross validation tend to give even better results as shown in table 2. Nevertheless, the overall results are promising for taking on any challenging dataset using the EPSC.

Initially it was supposed that an age balanced dataset would improve the overall classification, and therefore subsets of the AgeDB and CASIA was created. Even subspaces for each age group was created and tested. However, no great improvement was noticed. In the future, it should be tested whether, the approach by Gong et al. [GLJ20] could be used by dividing the learning data into groups, creating a subspace for each race and sex. Nonetheless, several of the datasets used contain a mix of races already, and it seem to work very well anyway.

Interestingly, the UTKFace dataset has the lowest accuracies when it comes to testing, but often yields the best accuracy when it is used as the set for learning. One reason for it being hard to classify, is that it contains quite many small children. They are hard to tell, even for humans, whether they are boys or girls.

The second hardest set to classify was CelebA and similarly it is the second best for learning, with one exception and that is when classifying on the NSD. Nevertheless, it seems like curating a set, requiring different age groups, did not help much. Better is to use a set with a great variation from the start. Also, even if CelebA is much larger than UTKFace, the latter performed better, perhaps because it is more gender balanced.

10 CONCLUSION

The proposed improvements to the EPSC, including how to learn projection depth and the improved damp-

ening function, makes EPSC reliable and stable for sex classification of face images. The face feature vectors from face recognition software, do include geometric information that can be used to determine sex and will not be affected by outer cues, such as facial hair, makeup and hair-style, which might be picked up by CNN based approaches, since they capture whatever is visible in the images. The results show close to state of the art performance, and even precede many of the algorithms published.

11 ACKNOWLEDGMENTS

This work has been partially supported by the Swedish Research Council (Dnr 2020-04652; Dnr 2022-02056) in the projects *The City's Faces. Visual culture and social structure in Stockholm 1880-1930* and *The International Centre for Evidence-Based Criminal Law (EB-CRIME)*. The computations were performed on resources provided by SNIC through UPPMAX under project SNIC 2021/22-918.

12 REFERENCES

[ADRS*19] Arrieta A. B., D'iaz-Rodr'iguez N., Ser J. D., Bennetot A., Tabik S., Barbado A., Garc'ia S., Gil-L'opez S., Molina D., Benjamins R., Chatila R., Herrera F.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *ArXiv abs/1910.10045* (2019).

[AHM*18] Abbas H., Hicks Y., Marshall D., Zhurov A., Richmond S.: A 3d morphometric perspective for facial gender analysis and classification using geodesic path curvature features. *Computational Visual Media 4* (01 2018), 1–16.

[AMVR*19] Acien A., Morales A., Vera-Rodriguez R., Bartolome I., Fierrez J.: Measuring the gender and ethnicity bias in deep models for face recognition. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (Cham, 2019), Vera-Rodriguez R., Fierrez J., Morales A., (Eds.), Springer International Publishing, pp. 584–593.

[BBD93] Burton A. M., Bruce V., Dench N.: What's the difference between men and women? evidence from facial measurement. *Perception 22*, 2 (1993), 153–176. PMID: 8474841.

[BBH*93] Bruce V., Burton A. M., Hanna E., Healey P., Mason O., Coombes A., Fright R., Linney A.: Sex discrimination: How do we tell the difference between male and female faces? *Perception 22*, 2 (1993), 131–152. PMID: 8474840.

[CPC19] Carvalho D. V., Pereira E. M., Cardoso J. S.: Machine learning interpretability: A survey on methods and metrics. *Electronics 8*, 8 (Jul 2019), 832.

[DDD*23] Donckt J. V. D., Donckt J. V. D., Deprost E., Vandenbussche N., Rademaker M., Vandewiele G., Hoecke S. V.: Do not sleep on traditional machine learning. *Biomedical Signal Processing and Control 81* (mar 2023), 104429.

[DGXZ19] Deng J., Guo J., Xue N., Zafeiriou S.: Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 4690–4699.

[DGZ*19] Deng J., Guo J., Zhou Y., Yu J., Kotsia I., Zafeiriou S.: Retinaface: Single-stage dense face localisation in the wild, 2019.

[EEH14] Eidinger E., Enbar R., Hassner T.: Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security 9*, 12 (2014), 2170–2179.

[GLJ20] Gong S., Liu X., Jain A. K.: Mitigating face recognition bias via group adaptive classifier, 2020.

[GLS90] Golomb B. A., Lawrence D. T., Sejnowski T. J.: Sexnet: A neural network identifies sex from human faces. In *NIPS* (1990).

[GM14] Gilani S. Z., Mian A.: Perceptual differences between men and women: A 3d facial morphometric perspective. In *2014 22nd International Conference on Pattern Recognition* (2014), pp. 2413–2418.

[GRS*14] Gilani S. Z., Rooney K., Shafait F., Walters M., Mian A.: Geometric facial gender scoring: Objectivity of perception. *PLOS ONE 9*, 6 (06 2014), 1–12.

[GSC*19] Gunning D., Stefik M., Choi J., Miller T., Stumpf S., Yang G.-Z.: Xai—explainable artificial intelligence. *Science Robotics 4*, 37 (2019).

[Hai04] Haig D.: The inexorable rise of gender and the decline of sex: Social change in academic titles, 1945-2001. *Arch Sex Behavior 33* (2004), 87–96.

[Has22] Hast A.: Magnitude of semicircle tiles in fourier-space : A handcrafted feature descriptor for word recognition using embedded prototype subspace classifiers. *Journal of WSCG 30*, 1-2 (2022), 82–90.

[Hin22] Hinton G.: The forward-forward algorithm: Some preliminary investigations, 2022.

[HL20] Hast A., Lind M.: Ensembles and cascading of embedded prototype subspace classifiers. *Journal of WSCG 28*, 1/2 (2020), 89–95.

[HLV19] Hast A., Lind M., Vats E.: Embedded prototype subspace classification : A subspace learning framework. In *The 18th International Conference on Computer Analysis of Images and Patterns (CAIP)* (2019), Lecture Notes in Computer Science, pp. 581–592.

[HRBLM07] Huang G. B., Ramesh M., Berg T., Learned-Miller E.: *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[HRGL05] Hoss R. A., Ramsey J. L., Griffin A. M., Langlois J. H.: The role of facial attractiveness and facial masculinity/femininity in sex classification of faces. *Perception 34*, 12 (2005), 1459–1474. PMID: 16457167.

[HV21] Hast A., Vats E.: Word recognition using embedded prototype subspace classifiers on a new imbalanced dataset. *Journal of WSCG 29*, 1-2 (2021), 39–47.

[Ins23] InsightFace: Insightface. https://insightface.ai, 2023. Accessed: 2023-02-30.

[KLR*77] Kohonen T., Lehtiö P., Rovamo J., Hyvärinen J., Bry K., Vainio L.: A principle of neural associative memory. *Neuroscience 2*, 6 (1977), 1065 – 1076.

[KO76] Kohonen T., Oja E.: Fast adaptive formation of orthogonalizing filters and associative memory in recurrent networks of neuron-like elements. *Biological Cybernetics 21*, 2 (Jun 1976), 85–95.

[Koh82] Kohonen T.: Self-organized formation of topologically correct feature maps. *Biological Cybernetics 43*, 1 (Jan. 1982), 59–69.

[Koh95] Kohavi R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2* (San Francisco, CA, USA, 1995), IJCAI'95, Morgan Kaufmann Publishers Inc., pp. 1137–1143.

[Kri19] Krishnan M.: Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology* (2019).

[KRMV76] Kohonen T., Reuhkala E., Mäkisara K., Vainio L.: Associative recall of images. *Biological Cybernetics 22*, 3 (Sep 1976), 159–168.

[KW96] Kohavi R., Wolpert D.: Bias plus variance decomposition for zero-one loss functions. In *Proceedings of the Thirteenth In-*

ternational Conference on International Conference on Machine Learning (San Francisco, CA, USA, 1996), ICML'96, Morgan Kaufmann Publishers Inc., pp. 275–283.

[Laa07] Laaksonen J.: *Subspace classifiers in recognition of handwritten digits*. G4 monografiaväitöskirja, Helsinki University of Technology, 1997-05-07.

[LH15] Levi G., Hassner T.: Age and gender classification using convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2015), pp. 34–42.

[LLWT15] Liu Z., Luo P., Wang X., Tang X.: Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)* (December 2015), pp. 3730–3738.

[LM14] Learned-Miller G. B. H. E.: *Labeled Faces in the Wild: Updates and New Reporting Procedures*. Tech. Rep. UM-CS-2014-003, University of Massachusetts, Amherst, May 2014.

[LWY*17] Liu W., Wen Y., Yu Z., Li M., Raj B., Song L.: Sphereface: Deep hypersphere embedding for face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA, USA, jul 2017), IEEE Computer Society, pp. 6738–6746.

[MH08] Maaten L. v. d., Hinton G.: Visualizing data using t-sne. *Journal of machine learning research 9*, Nov (2008), 2579–2605.

[MH18] McInnes L., Healy J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints* (Feb. 2018).

[MPS*17] Moschoglou S., Papaioannou A., Sagonas C., Deng J., Kotsia I., Zafeiriou S.: Agedb: The first manually collected, in-the-wild age database. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2017), pp. 1997–2005.

[MR08] Makinen E., Raisamo R.: Evaluation of gender classification methods with automatically detected and aligned faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 3 (2008), 541–547.

[NLSS04] Nordstrøm M. M., Larsen M., Sierakowski J., Stegmann M. B.: *The IMM Face Database - An Annotated Dataset of 240 Face Images*. Tech. rep., Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, may 2004.

[oE23] of Europe C.: Sex and gender. https://www.coe.int/en/web/gender-matters/sex-and-gender, 2023. Accessed: 2023-02-23.

[OK83] Oja E., Kuusela M.: The alsm algorithm - an improved subspace method of classification. *Pattern Recognition 16*, 4 (1983), 421–427.

[OK88] Oja E., Kohonen T.: The subspace learning algorithm as a formalism for pattern recognition and neural networks. In *IEEE 1988 International Conference on Neural Networks* (July 1988), vol. 1, pp. 277–284.

[OPD96] O'Toole A., Peterson J., Deffenbacher K.: An 'other-race effect' for categorizing faces by sex. *Perception 25* (02 1996), 669–76.

[PWHR98] Phillips P., Wechsler H., Huang J., Rauss P. J.: The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing 16*, 5 (1998), 295–306.

[RPC16] Ranjan R., Patel V., Chellappa R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence PP* (03 2016).

[SBLM17] Samek W., Binder A., Lapuschkin S., MÃ¼ller K.-R.: Understanding and comparing deep neural networks for age and gender classification. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)* (2017), pp. 1629–1638.

[Sha18] Shapshak P.: Artificial intelligence and brain. *Bioinformation 14*, 1 (2018), 38.

[SHIA21] Sumi T. A., Hossain M. S., Islam R. U., Andersson K.: Human gender detection from facial images using convolution neural network. In *Applied Intelligence and Informatics* (Cham, 2021), Mahmud M., Kaiser M. S., Kasabov N., Iftekharuddin K., Zhong N., (Eds.), Springer International Publishing, pp. 188–203.

[SJ19] Shi Y., Jain A.: Probabilistic face embeddings. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 6901–6910.

[SKP15] Schroff F., Kalenichenko D., Philbin J.: Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 815–823.

[TYRW14] Taigman Y., Yang M., Ranzato M., Wolf L.: Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 1701–1708.

[VJ01] Viola P., Jones M.: Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001* (Dec 2001), vol. 1, pp. I–I.

[WDH*19] Wang M., Deng W., Hu J., Tao X., Huang Y.: Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Los Alamitos, CA, USA, nov 2019), IEEE Computer Society, pp. 692–702.

[WLK*67] Watanabe W., Lambert P. F., Kulikowski C. A., Buxto J. L., Walker R.: Evaluation and selection of variables in pattern recognition. In *Computer and Information Sciences* (1967), Tou J., (Ed.), vol. 2, New York: Academic Press, pp. 91–122.

[WP73] Watanabe S., Pakvasa N.: Subspace method in pattern recognition. In *1st Int. J. Conference on Pattern Recognition, Washington DC* (1973), pp. 25–32.

[WWZ*18] Wang H., Wang Y., Zhou Z., Ji X., Gong D., Zhou J., Li Z., Liu W.: Cosface: Large margin cosine loss for deep face recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 5265–5274.

[WZLQ16] Wen Y., Zhang K., Li Z., Qiao Y.: A discriminative feature learning approach for deep face recognition. In *Computer Vision – ECCV 2016* (Cham, 2016), Leibe B., Matas J., Sebe N., Welling M., (Eds.), Springer International Publishing, pp. 499–515.

[YLLL14] Yi D., Lei Z., Liao S., Li S. Z.: Learning face representation from scratch, 2014.

[ZYQ17] Zhifei Z., Yang S., Qi H.: Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), IEEE.