Chapter 1

Blinding with Linear Clustering Removal

Face recognition algorithms work with embedding spaces. They map images of persons into the embedding space such that images of the same person are close to each other in the embedding space. This work investigates discriminatory dimensions in face recognition algorithms. For a given discriminatory dimension, the data can be grouped into clusters. A blinding procedure is proposed to remove the information related to the separation of these clusters. The procedure is a linear operation in the embedding space and uses the following steps:

1. Compute centers of clusters defined by the discriminatory dimension.
2. Use a one-vs-rest (OvR) Ansatz to calculate the directions of discrimination of each cluster relative to the other clusters.
3. Apply singalur value decomposition (SVD) on the directions of discrimination to find an orthonormal basis spanning the "discriminatory subspace".
4. Remove projections onto the "discriminatory subspace" from the embedding vectors. This results in embedding vectors which are orthogonal to the directions of discrimination.

After outlining the method, cluster visualization, awareness and face recognition rates are investigated before and after the blinding procedure.

1.1 The math behind

In the following we look at the discriminatory dimension of race. We work with the commonly used racial faces in-the-wild (RFW) data set which groups faces into $K = 4$ ethnic clusters Caucasian, African, Asian and Indian. I consider a VGG2 model where the embedding space has $N_e = 128$ dimensions. The procedure outlined above operates on the empedding vectors $\boldsymbol{x}_i$ where i denotes the sample. Associated to each sample is a cluster label $k \in \{1, \ldots, K\}$. As stated, the goal is to remove the directions in the embedding space which separate the ethnic clusters. As a first step, we define the centers of each cluster by the average

$$\bar{\boldsymbol{x}}_k = \frac{1}{n_k} \sum_{i \in C_k} \boldsymbol{x}_i, \tag{1.1}$$

where C_k is the set of embedding vectors associated with cluster k and n_k is the corresponding size. Following a one-vs-rest (OvR) approach, the normalized direction of discrimination of each cluster k to the other clusters is given by the vectors

$$\boldsymbol{u}_k = \frac{\boldsymbol{v}_k}{\|\boldsymbol{v}_k\|} \quad \text{with} \quad \boldsymbol{v}_k = \bar{\boldsymbol{x}}_k - \frac{1}{K-1} \sum_{k' \neq k} \bar{\boldsymbol{x}}_{k'}, \tag{1.2}$$

where K is the number of clusters. Fig. 1.1 shows the projections onto the vectors $\boldsymbol{u}_k$. As it may be expected from the construction, each direction nicely separates the corresponding cluster from the others. These means, that different ethnic groups are literally located in different corners of the empedding space. By construction, the vectors $\boldsymbol{u}_k$ are not linearly independent, but span a subspace of rank $K-1$. This can by verified by applying a singular value decomposition (SVD) on the matrix $U = [\boldsymbol{u}_1 \ldots \boldsymbol{u}_K]$. SVD also provides a orthonormal basis $B = [\boldsymbol{e}_1 \ldots \boldsymbol{e}_{K-1}]$ of the corresponding subspace. The final step is to remove the projections onto this subspace by

$$\boldsymbol{x}_i^b = \boldsymbol{x}_i - \sum_{j=1}^{K-1} (\boldsymbol{x}_i \cdot \boldsymbol{e}_j)\, \boldsymbol{e}_j, \tag{1.3}$$

where $(\boldsymbol{x}_i \cdot \boldsymbol{e}_j)$ is the dot (or scalar) product. Eq. (1.3) yields new embedding vectors $\boldsymbol{x}_i^b$ with the same shape as the original ones. The upper index b stands for *blinded* inspired by the fact that some information with regard to the discriminatory dimension has been removed. Note that the new embeddings depend linearly on the original ones.

1.2 Awareness

Awareness is the ability of the model to discriminate between different clusters of the discriminatory dimensions, being the ethnic label in present case. This ability obviously depends on the trained model at hand. Here I benchmark the performance to predict the ethnic labels by the *aware* embeddings $\boldsymbol{x}_i$ and the *blinded* ones $\boldsymbol{x}_i^b$. A train/test split of two thirds/one third was used. The accuracy of different classifiers is shown in Tab. 1.1. Not surprisingly, linear classifiers are unable to predict the race for the blinded embeddings. Nearest neighbor approaches still work reasonably. More advanced non-linear classifiers

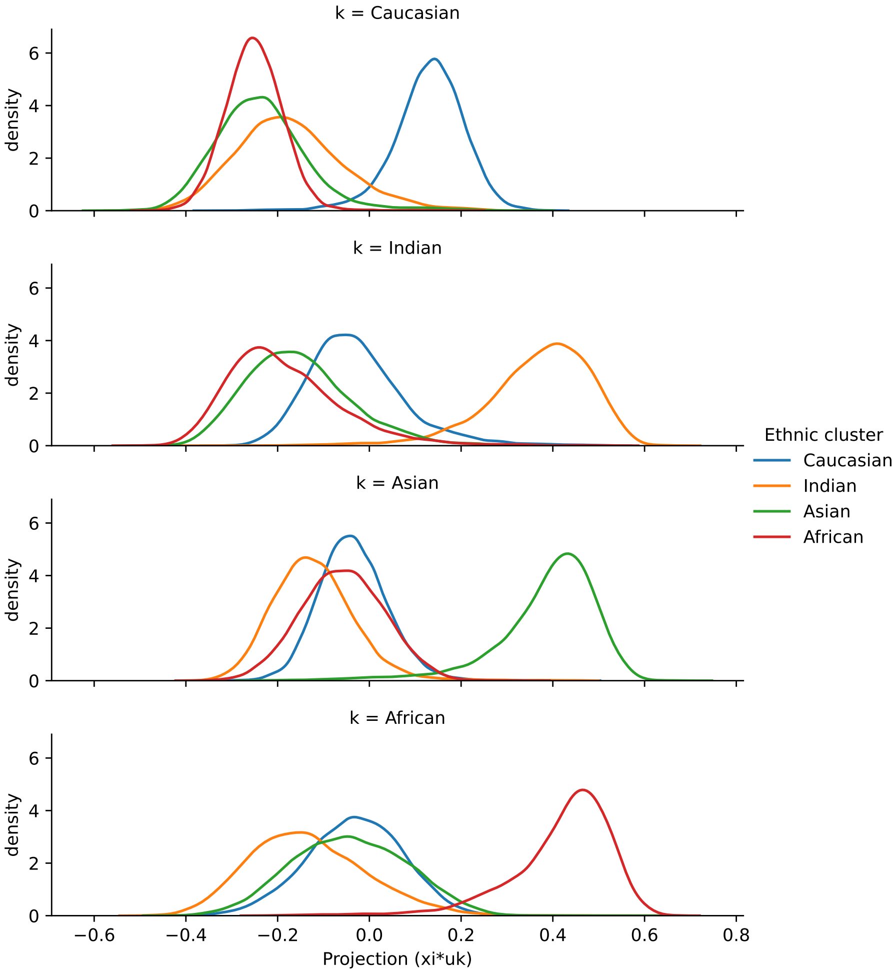

Figure 1.1: Kernel density estimation (KDE) plot of the projections $(\boldsymbol{x}_i \cdot \boldsymbol{u}_k)$. The normalized directions $\boldsymbol{u}_k$ represent the discriminatory directions which separate each cluster from the others.

Model	aware	blinded
Logistic regression	96%	21%
Linear SVM	96%	26%
Nearest neighbor	92%	57%
5 Nearest neighbor	94%	62%
NN with 1 hidden layer (100 nodes), relu	96%	70%
NN with 2 hidden layer (100 nodes each), relu	96%	85%

Table 1.1: Subset accuracy of various classifiers predicting the ethnic labels based on *aware* and *blinded* emdeddings. A train/test split of two thirds/one third was used.

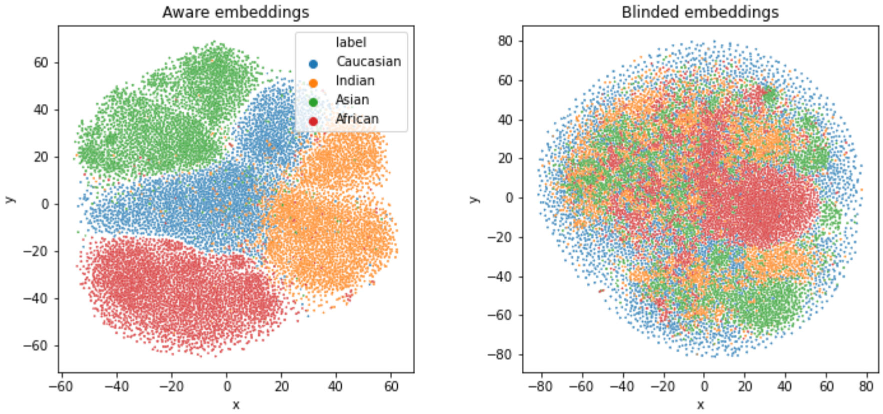

Figure 1.2: tSNE plots of the two embeddings

such as neural networks perform well. The clustering displayed by the corresponding t-SNE plots in Fig. 1.2 is in line with these findings. In the blinded case, clusters can't be separated be a single straight line. However, the data still displays groups defined by ethnical labels. Interestingly, there are clear differences between the groups. Africans are grouped in the center, Caucasian encircle the cloud and Indians/Asian are scattered inbetween.

1.3 Cluster scores

Cluster scores give a measure of clustering. They are calculated for both embeddings in Tab. 1.2. The Silhoutte cluster score gives a measure between 0 and 1 indicating how well the data is clustered. The Silhoutte cluster score of the aware embeddings is only 0.063 - basically indicating the absence of clustering although Fig. 1.1 and Fig. 1.2 show nice clustering for the aware embeddings. This counter-intuitive finding is due to the high dimensionality. Both figures

Cluster score	aware	blinded
Silhouette score	0.063	-0.014
Calinski-Harabasz score	1814	0
Davies-Bouldin score	3.8	2.8x10^6

Table 1.2: Cluster scores for *aware* and *blinded* emdeddings.

N_e	128 aware	128 blinded	256 aware	256 blinded	2048 aware	2048 blinded
Total	86%	88%	86%	88%	83%	84%
Caucasian	91%	92%	91%	92%	89%	89%
Indian	86%	88%	86%	88%	85%	85%
Asian	84%	86%	84%	87%	82%	82%
African	84%	86%	84%	85%	76%	79%

Table 1.3: Face recognition rates of the RFW dataset for *aware* and *blinded* emdeddings and for different Senet models indicated by the embedding size N_e. The threshold was optimized for each case (corresponding to a column) with respect to the total dataset.

show projections into lower dimensions and therefore reflect only a marginal part of the information. This is confirmed by the fact that the total variance of the blinded embeddings is still 84% of the original variance.

1.4 Face recognition rates and bias

Face recognition rates and bias are evaluated with the RFW data set. The RFW dataset provides image (i.e. embedding) pairs corresponding to the same or to different persons. The resulting task is a binary classification of the pairs into "same" and "different". The recognition rate is the accuracy of the corresponding classification. The feature used for the classification is the pair distance in the embedding space. Here we use the cosine distance:

$$d_{ij} = 1 - \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_j}{\|\boldsymbol{x}_i\| \, \|\boldsymbol{x}_j\|} \tag{1.4}$$

The face recognition rates calculated in this way are shown in Tab. 1.3. The table includes further Senet models with $N_e = 256, 2048$. Surprisingly, the performance increases for the blinded embeddings by about 2% for all clusters. Bias is slightly removed. Fig. 1.3 gives further insights by showing the distribution of the cosine distances for "same" and "different" pairs. The Caucasians are special in that their distributions are not significantly altered. Note that the blinding procedure leads to a better alignment of the thresholds.

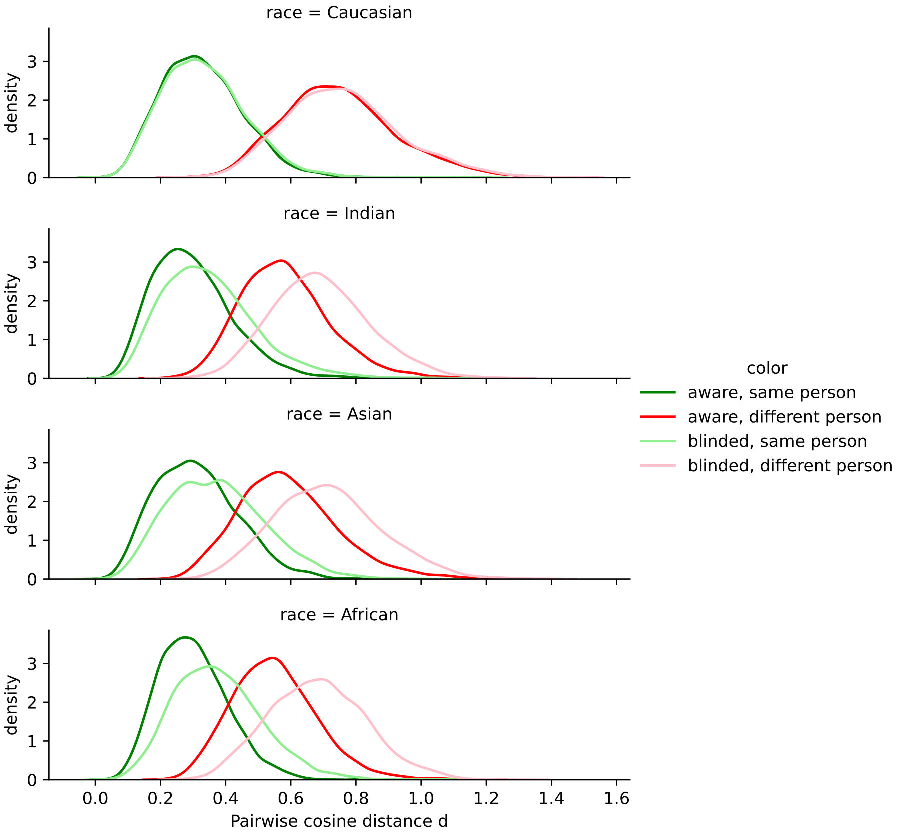

Figure 1.3: Kernel density estimation (KDE) plot of pairwise cosine distances for same persons (green) and different persons (red) and for *aware* (darker color) and *blinded* (lighter color) emdeddings.

1.5 Discussion

Most of the findings with regard to the blinding procedures are not surprising. The proposed procedure removes linear separability with the effect that linear classifiers can't distinguish between ethnic clusters after blinding in line with the cluster scores. The fact that the ethnic clusters are well separated in the first place shows that the considered model clearly distinguishes the ethnic groups. The blinding removes this separation. The surprise comes at the end. The removal of this information actually improves overall performance and slightly reduces bias. The performance improvement through this simple linear approach is very surprising. A reason might be that there are a substantial differences between the data used for training of the model and the RFW data used for present benchmark. At this stage it is not clear, whether this performance improvement is accidental or generalizes to other models and other discriminatory dimensions.