The aging effect across human races

Do they preserve the same way?

Group 12: s220492 & s213237 & s212487 & s182105

Computational Data Analysis, Spring 2022

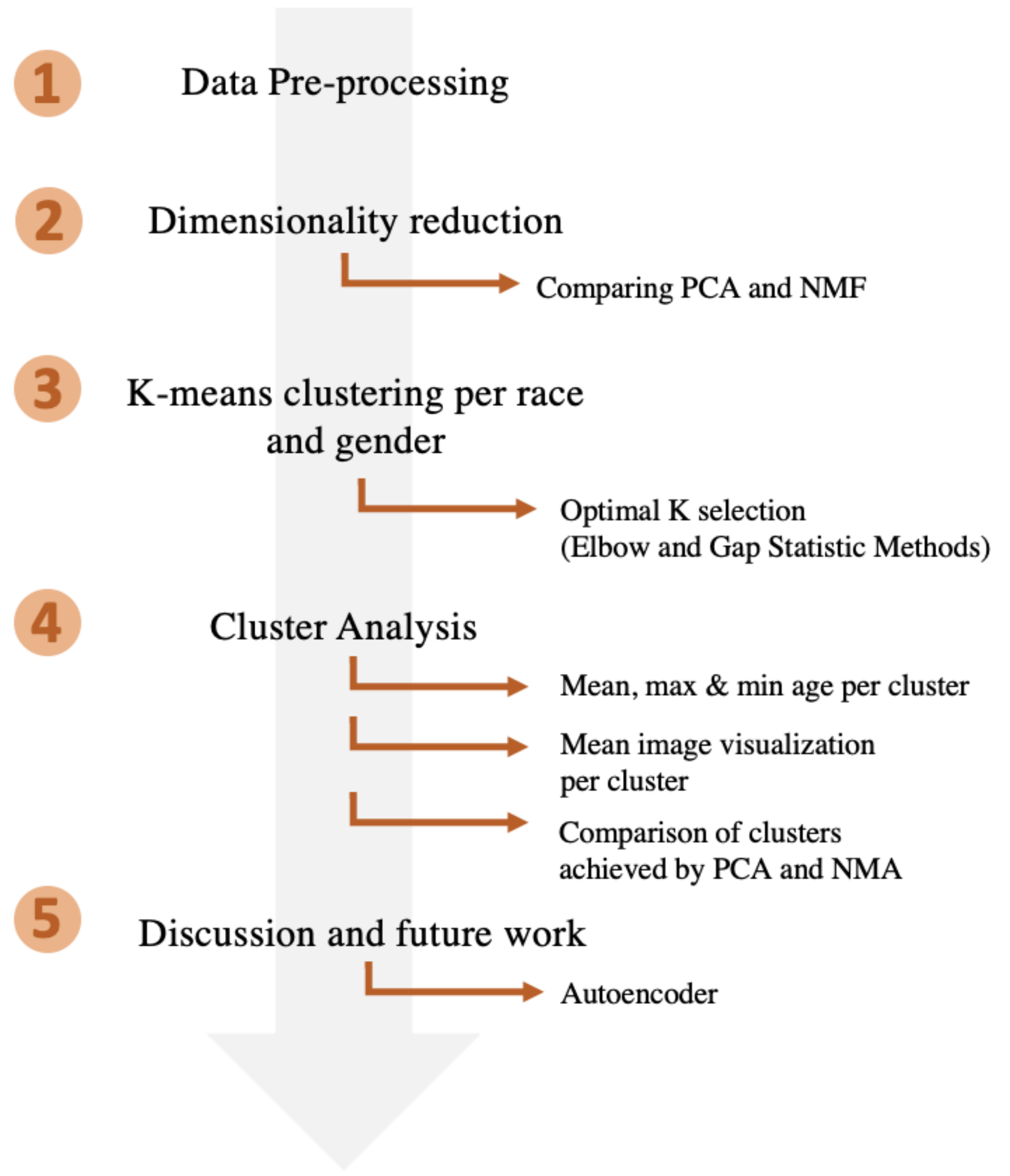
DTU Danmarks Tekniske Universitet

Introduction

Image analysis on human data is a hot topic within the Computer Vision area. Furthermore, narrowing the subject down to the analysis of human faces sparks special curiosity among the global population. Many questions arise surrounding this matter; *Do certain facial features clearly determine the age of a person? Can we extract which ones? Which races age better?*

Main Objectives

The motivation for this project follows the questions posed above. Our aim is to (1) extract the most relevant features per race and gender groups using Principal Component Analysis (PCA) and Non-negative Matrix Normalisation (NMF) and (2) **explore the age clusters that an unsupervised learning algorithm can make across these subgroups by considering the extracted face features.**



1 Data Pre-processing

2 Dimensionality reduction

Comparing PCA and NMF

3 K-means clustering per race and gender

Optimal K selection (Elbow and Gap Statistic Methods)

4 Cluster Analysis

Mean, max & min age per cluster

Mean image visualization per cluster

Comparison of clusters achieved by PCA and NMA

5 Discussion and future work

Autoencoder

Figure 1: Pipeline followed to attain our goals.

Data and preprocessing

The image data given by **UTK face dataset** was used, with a total number of 23705 different faces of 200x200 [1]. The RGB images were converted to a grayscale version as the color was not relevant for our problem statement, thus reducing data dimensionality. The face images were then vectorized to compact all data in a single array (n=23705,p=40000). For next analyses, we filter data by race and gender, ending up with 10 subgroups (5 races, per 2 genders each).

Feature Extraction: PCA vs NMF

To achieve dimension reduction of the data, two methods were implemented in parallel for better comparison of results: (1) **Principal Component analysis (PCA)** and (2) **Non-negative Matrix Factorization (NMF)**. While both pursue the goal of extracting the number of components that are most meaningful, there are still some subtle differences. While PCA gives "generic" faces ordered by how well they capture the original one, NMF creates a sparse representation, using less components to describe the image but retaining face features such as eyes, nose, mouth and eyebrows which are the main characteristics of the data. Moreover, for the PCA case, to enhance the extraction of meaningful features and avoid the clustering in terms of the image background, an ellipsoid mask was created in the centre of each image to be able to apply PCA to just those pixels inside the mask.

	PCA / NMF Feature Extraction
1	Mean img substraction for each subgroup
2	Determination of number of components
3	PCA/NMF implementation from Sklearn library
4	Retrieval of the most meaningful features

Table 1: Feature extraction procedure for PCA and NMR

Data reduction was achieved by retaining a total number of 200 principal components for the PCA, which accounts for 96-98% of the explained variance for every subgroup. NMF has been implemented by using just 4 components. This dimensional reduction of data allowed a faster performance of the clustering algorithm.

K-Means Clustering

The K-Means Clustering was the unsupervised learning algorithm used to explore age clusters created by the selected features from PCA, and the groups with the components from NMR.

Optimal K

The selection of the optimal K (number of clusters) was the first step taken, applying both the elbow method and gap statistic. Both lead to the same results: an optimal K value of 4 for PCA data, and a k value of 5 for the NMF data.

Cluster Analysis

For a better understanding of cluster performance, the mean (original) image per cluster for each race was obtained for both men and women. Our results showed that PCA data did not perform as expected, probably due to the fact that with this dimension reduction, a 2-4% of the explained variance was lost, which might be related to the face age features we were looking for. As a result, the unsupervised clustering was performed between more generic faces, and thus leading to a poor quality of results. In contrast, NMF data performed better,in terms that the clustering actually distinguishes between age groups slightly better across the different races. This was the reason why we decided to move forward and present these results.

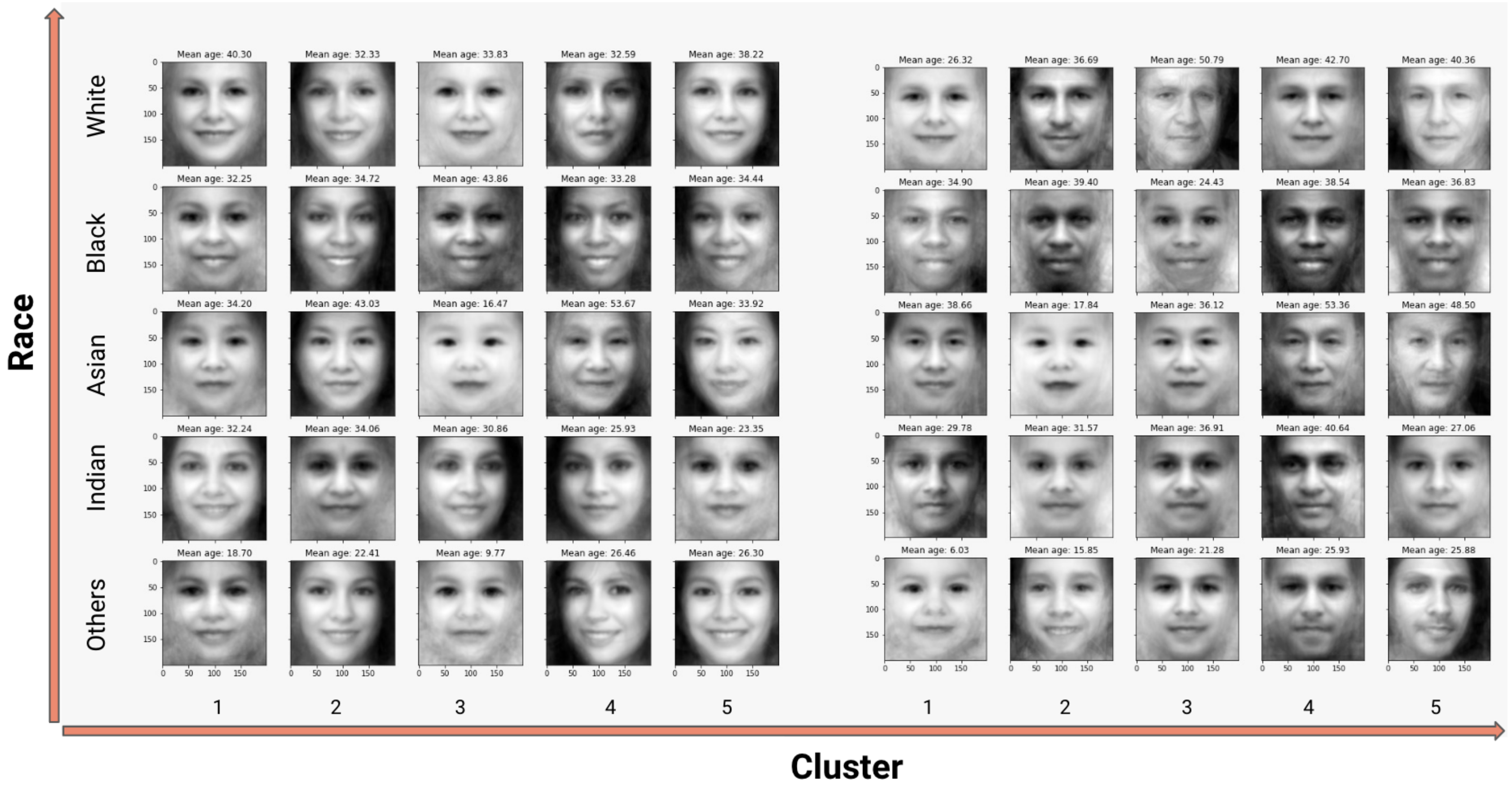


Figure 2: Mean of original images for each cluster per race and gender using features extracted by NMF, together with the mean age.

Discussion

In order to address the main question of interest (*which race and gender ages better?*), the focus needs to be put on what the race clusters show, and specially what is the mean age per cluster. Races showing more determining age features will lead to better clustering, and thus, show bigger differences with respect to the mean age per cluster.

With respect to men, it can be clearly seen that clusters are quite homogeneous and that the biggest differences can be observed between children and adults. However, Black and Indian races seem to age better than the others, due to clusters show less deviation in terms of the mean age. In other words, clustering is worse done due to less determining age features are found, meaning that they basically age better.

In regards to women, the differentiation between clusters is much more weaker, and in most of the races is hard to distinguish between children and adults. Face features seem less determining, meaning that in general terms it could be said that women age better than men. However, still some differences are observed between races, being the Black race the one that ages better.

Overall, it could be said that main differences between men and women might be due to the fact that men have a clear feature when they age called beard. Moreover, it can be concluded that clustering by age was noticeable, thus full-filling our main goal, but leaving room for further improvements.

Limitations and future work

Much could be done to enhance feature extraction related to age. The Auto-encoder was also considered, to perform a non-linear approach for the dimension reduction of the data, and see if it lead to a better outcome compared to the PCA and NMF linear methods. However, this technique is computationally expensive, and our current resources didn't allow to explore further.

Future work could investigate this method in deep, also adding more powerful unsupervised clustering algorithms enhanced with regularisation techniques.

References

[1] Utkface, 2017.

[2] elenageminiani. Nmf and image compression, 06 2020.

[3] S Joel Franklin. Pros cons of dimensionality reduction, datadriven-investor, 01 2020.