Project of Processing Big Data

Group 8:
Vasco Carneiro, 93359 and Tiago Miranda, 93416

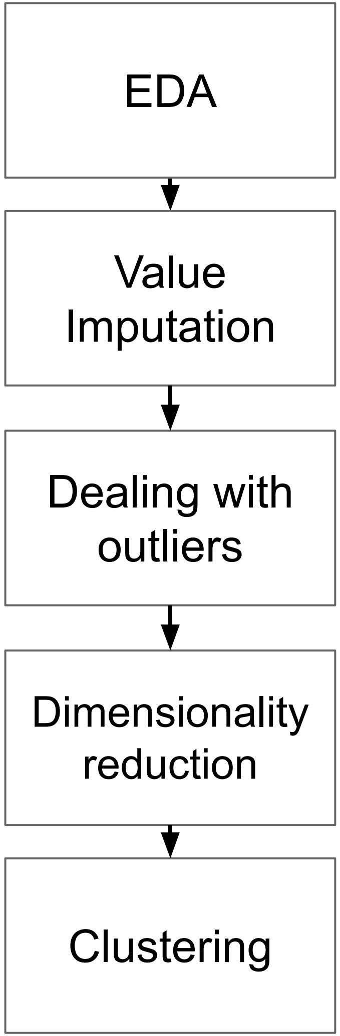
Introduction

We did it for 3 methods using the dauphine dataset.

We clustered the following datasets:

1. Skeletons
2. Image embeddings
3. Skeletons + images

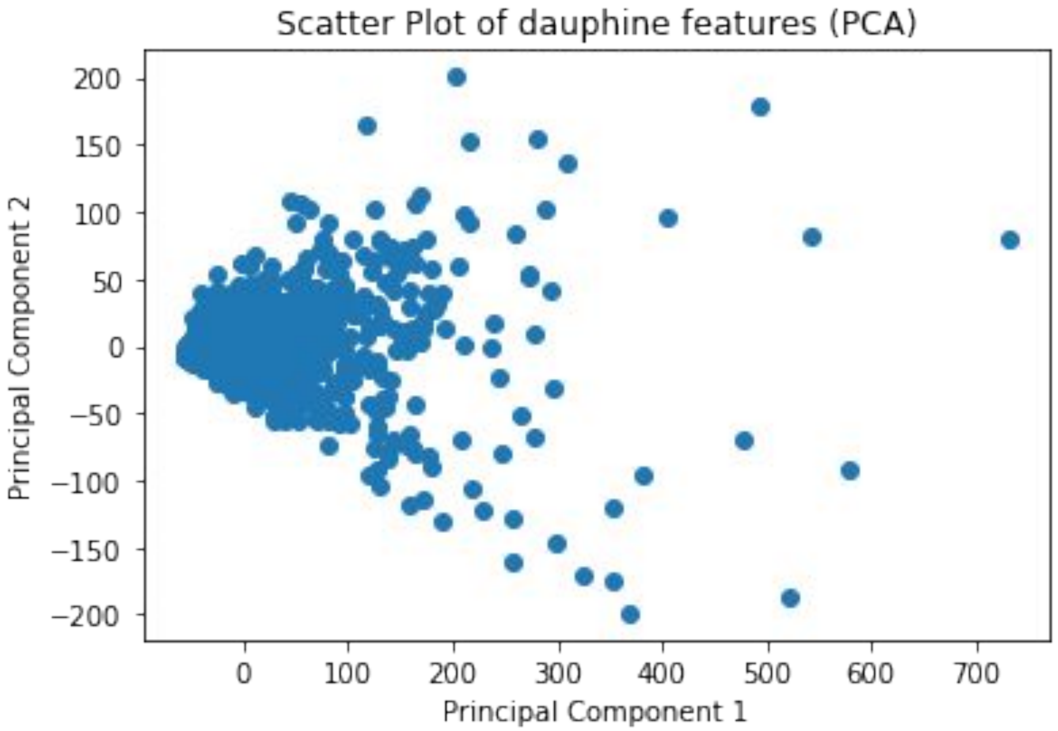
Pipeline


EDA
Value Imputation
Dealing with outliers
Dimensionality reduction
Clustering

EDA – Dauphine features

```
In [5]:  1  np.shape(dauphine_features)
Out[5]:  (2048, 10734)
```

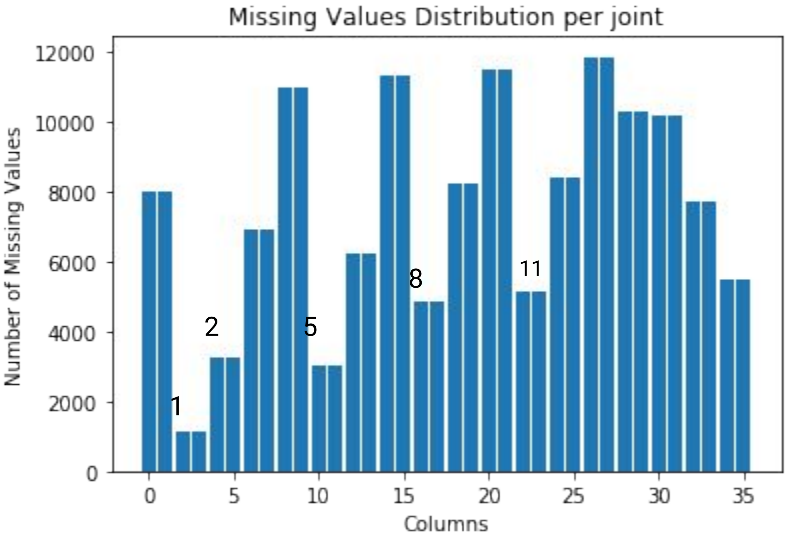
```
threshold for distant points= 400
percentage of distant points: 0.390625
Index: 67
Index: 124
Index: 138
Index: 520
Index: 945
Index: 1147
Index: 1222
Index: 1250
```

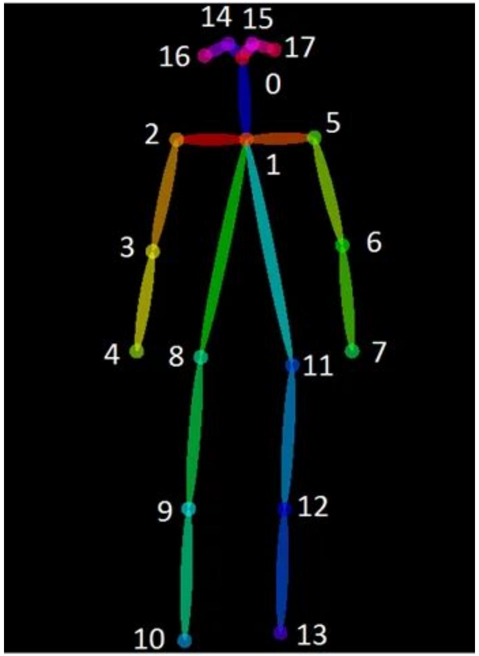


EDA – Dauphine Incomplete Skeletons

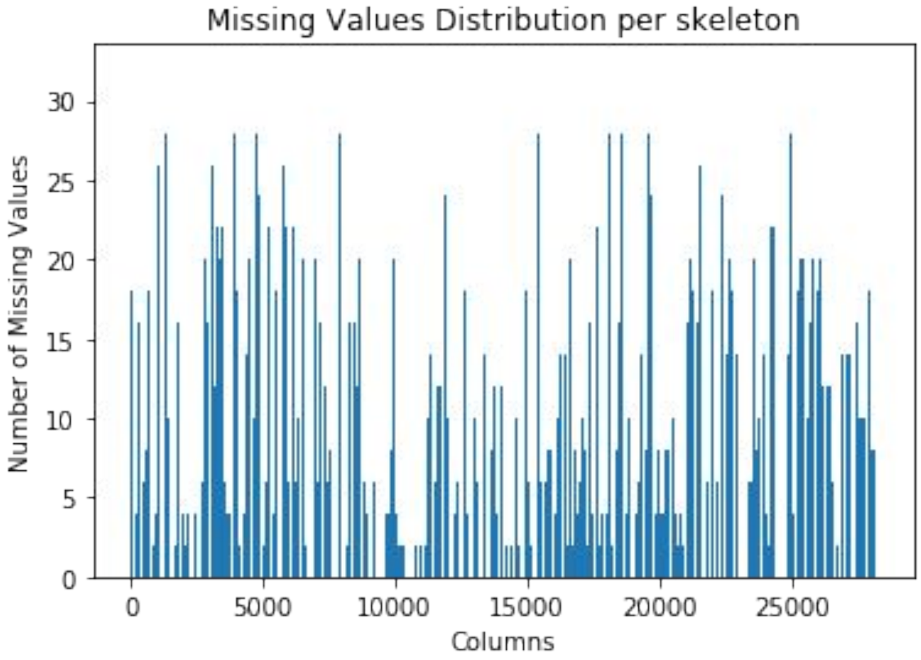
```
#without probabilities, only coordenates
np.shape(dauphine_skeletons)
```

(28232, 36)





EDA – Dauphine Incomplete Skeletons




```
#without probabilities, only coordenates
np.shape(dauphine_skeletons)
```

(28232, 36)

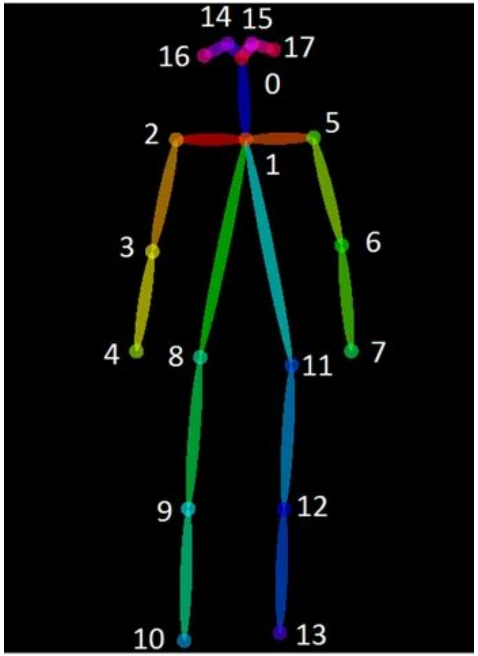
Missing value imputation

Methods tested:

- ❖ Mean value imputation
- ❖ MICE
- ❖ K-nearest neighbours
- ❖ GLRM usando h2o

Standardizing skeleton data

Openpose



$$z = (x - \mu) / \sigma$$

$$\bar{x}_{\text{openpose}} = \frac{x_1 + x_2 + x_5}{3}$$

$$\bar{y}_{\text{openpose}} = \frac{y_1 + y_2 + y_5}{3}$$

$$\bar{x}_{\text{otherpose}} = \frac{x_6 + x_7}{2}$$

$$\bar{y}_{\text{otherpose}} = \frac{y_6 + y_7}{2}$$

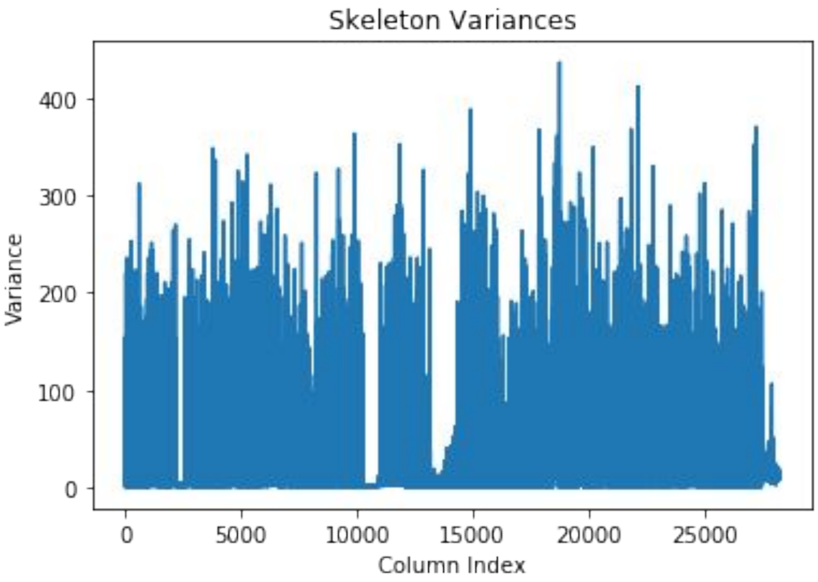
Otherpose

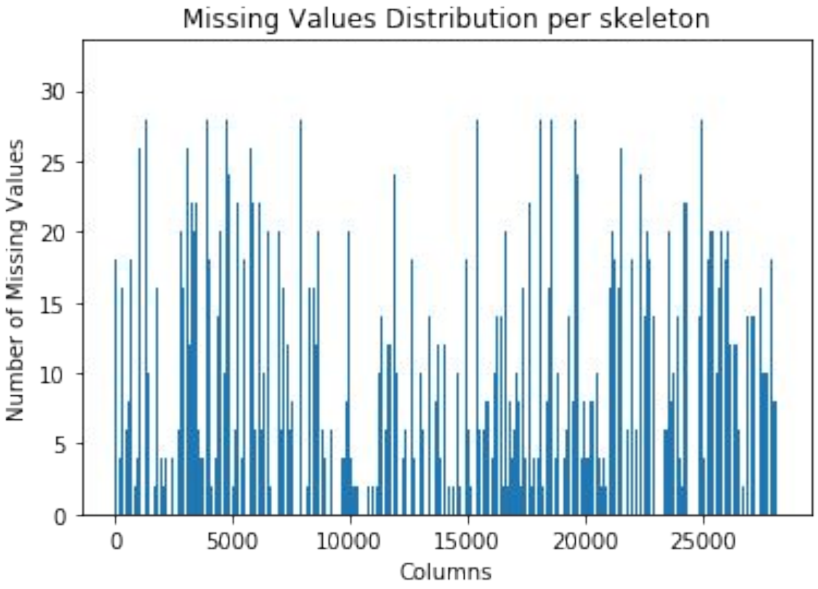
GLRM

$$m\left\{\overbrace{\begin{bmatrix} \quad & A & \quad \end{bmatrix}}^{n}\right. \approx m\left\{\overbrace{\begin{bmatrix} X \end{bmatrix}}^{k}\right. \left.\overbrace{\begin{bmatrix} \quad & Y & \quad \end{bmatrix}}^{n}\right\} k$$

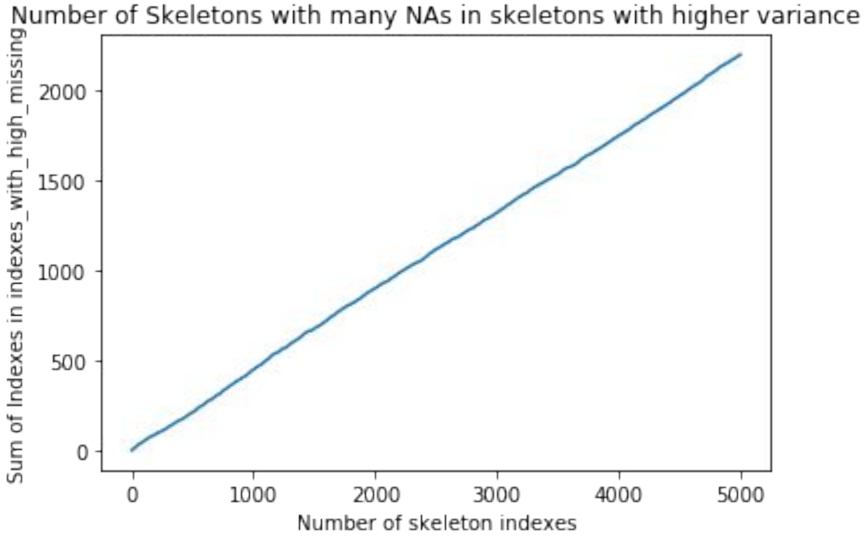
```python
# Define and train the GLRM model to impute missing values
glrm_model = H2OGeneralizedLowRankEstimator(k=20,
                                            loss="Quadratic",
                                            regularization_x="L1",
                                            regularization_y="L1",
                                            max_iterations=100,
                                            recover_svd=True)
```

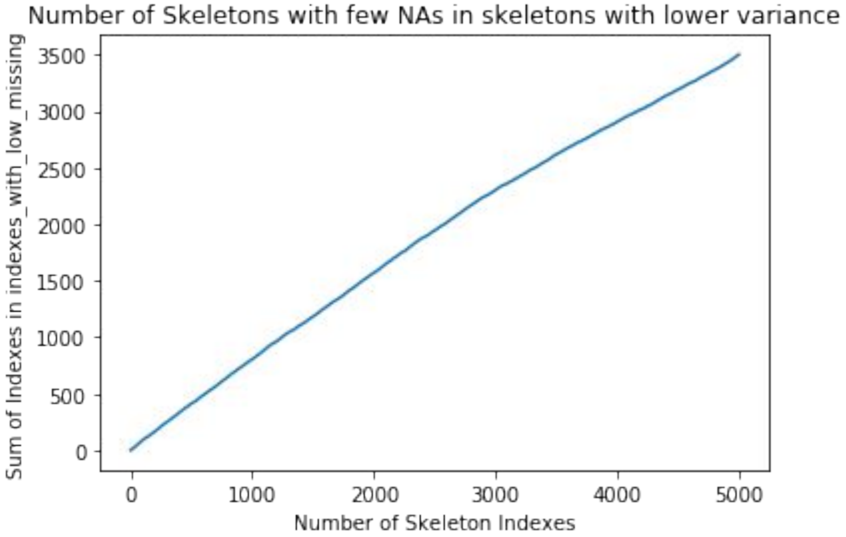
Dauphine Complete Skeletons



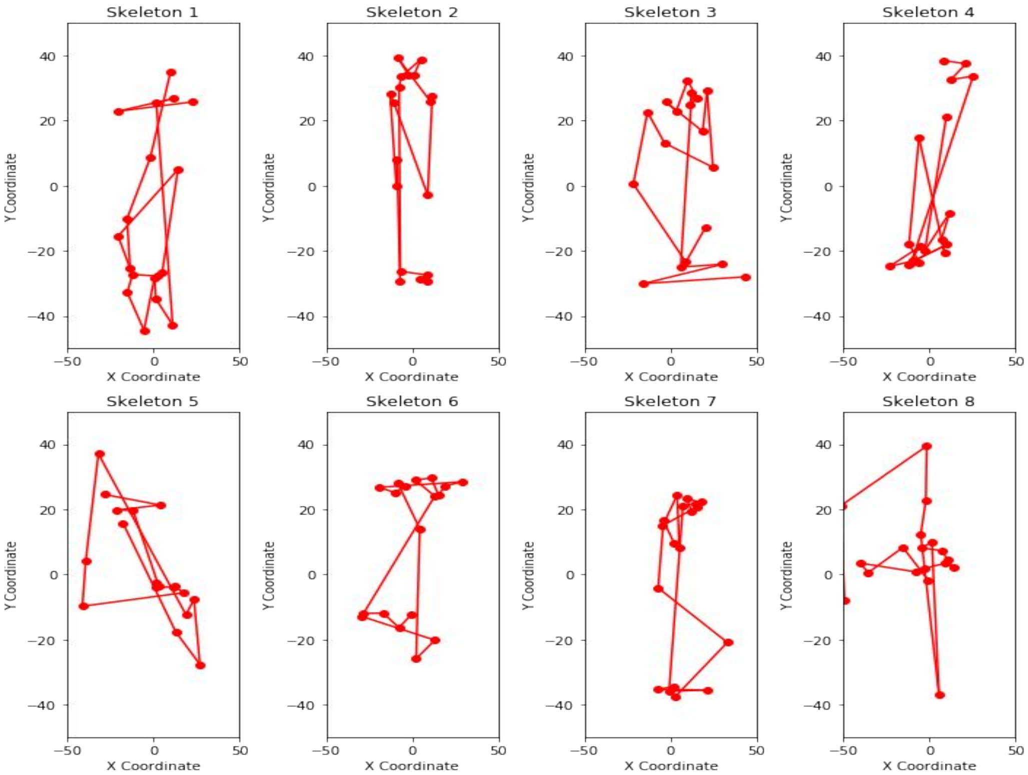


Dauphine Complete Skeletons

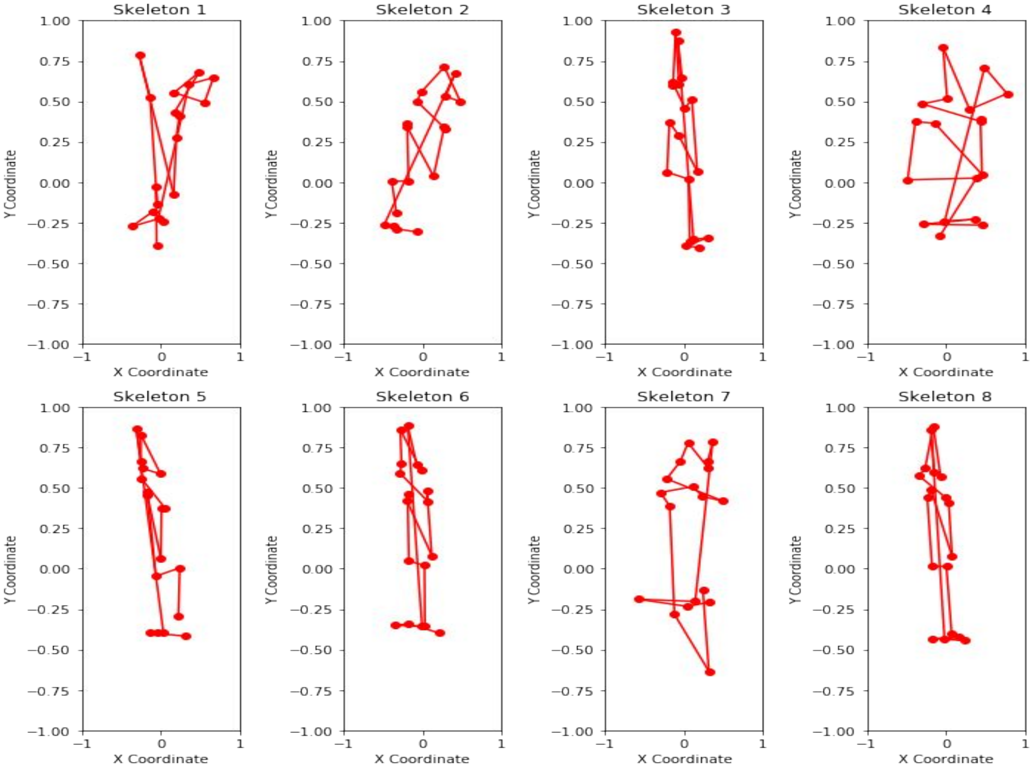




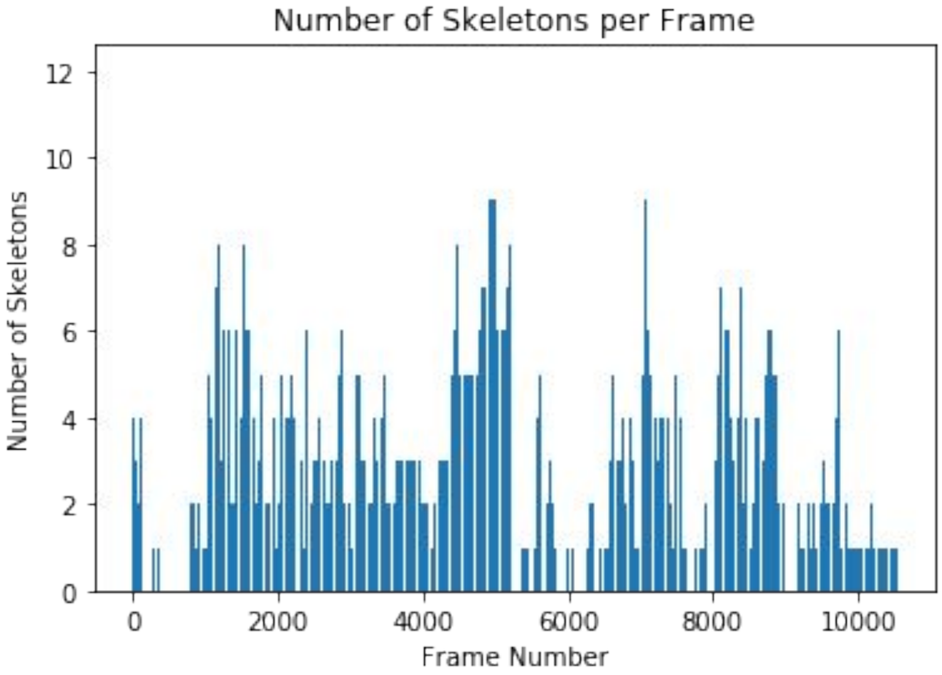
Dauphine Complete Skeletons



Dauphine Complete Skeletons



Dauphine Complete Skeletons



Detecting rank

- For skeleton and image embeddings data the method is the same
- Perform SVD and *r* corresponds to the number of components that explain at least 99.9% of cumulative explained variance

For skeleton the rank is 20.

For image embeddings is 95.

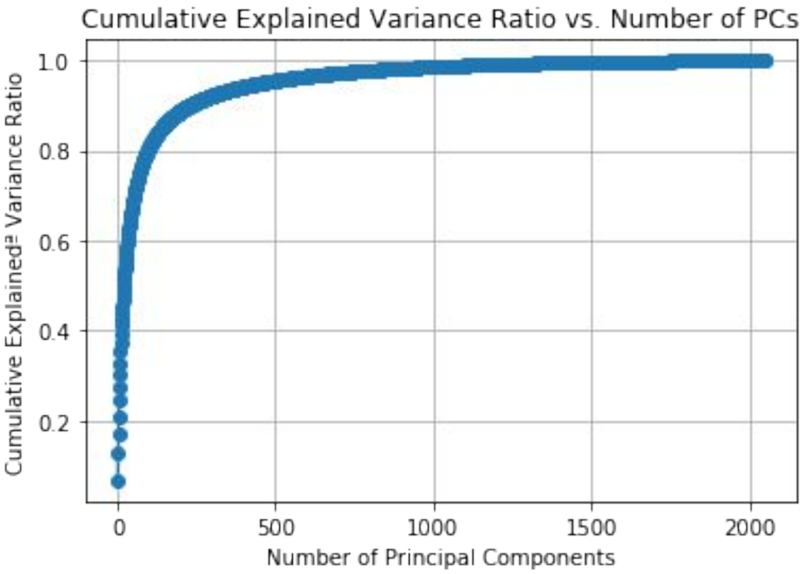
Removing outliers and reducing dimensionality

- For skeleton and image embeddings data the method is the same
- We calculate the distance to the Null space

Reducing dimensionality

- Take the SVD obtained and only retain the most important *r* components

Dimensionality Reduction – Features




```
Number of components necessary for  80.0 %% cumulative explained percentage is  95
Shape of reduced_data: (10734, 95)
(10734, 95)
```

Clustering – The applied methods and datasets

We performed the following clusterings:

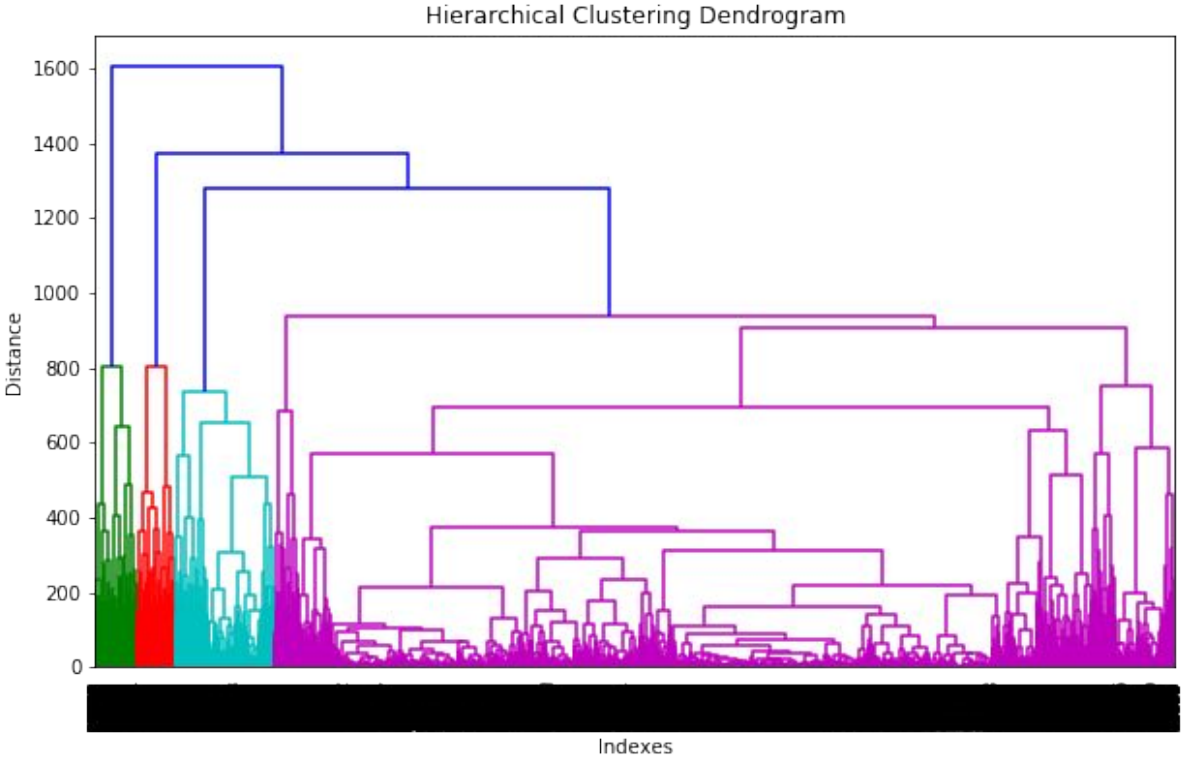
- ❖ K-Means

- ❖ Hierarchical Clustering:
  - ➢ Complete Linkage
  - ➢ Ward Method
  - ➢ Centroid Method

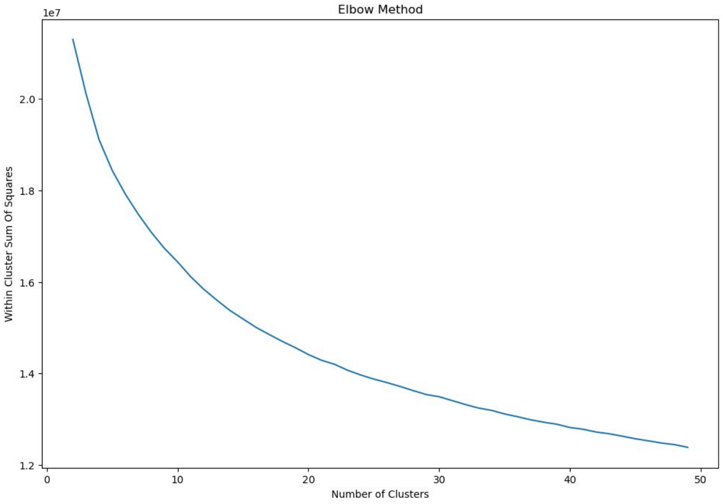
We clustered the following datasets:

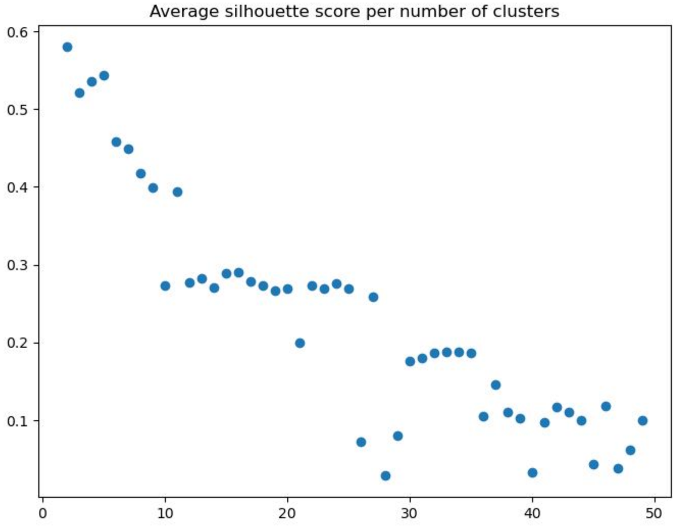
1. Skeletons
2. Image embeddings
3. Skeletons + images

1. Skeletons: Hierarchical Clustering – Ward Method



1. Skeletons – Number of clusters





Clustering visualization only for skeleton data

Cluster

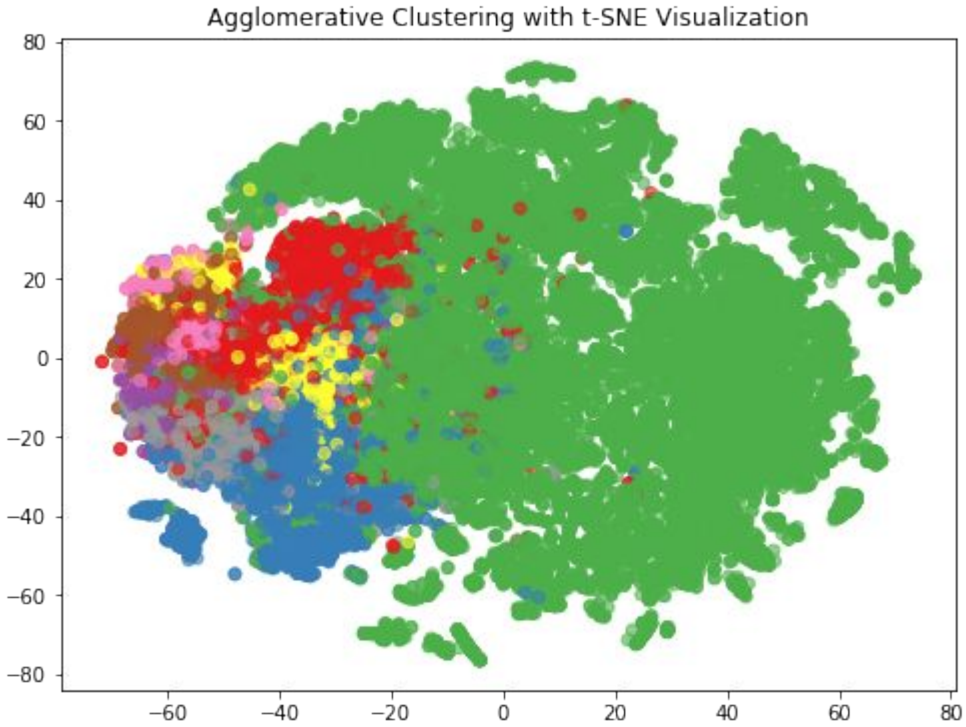
0

1

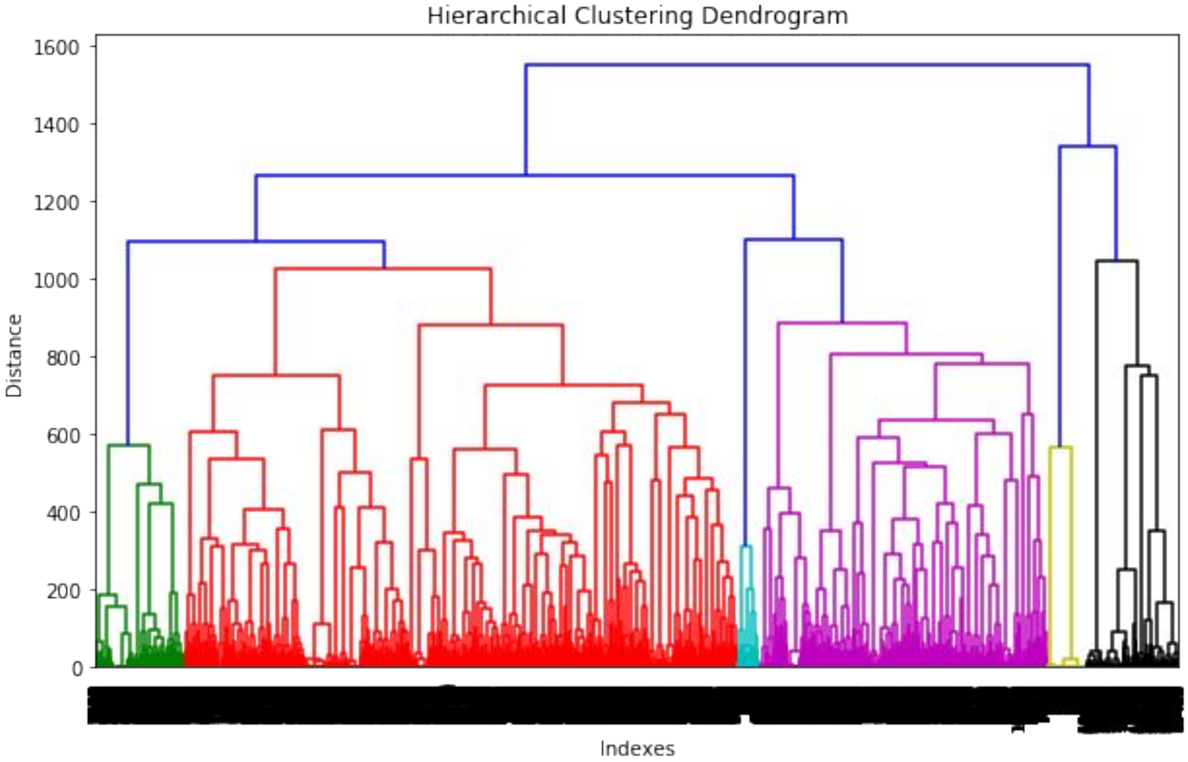
2

3

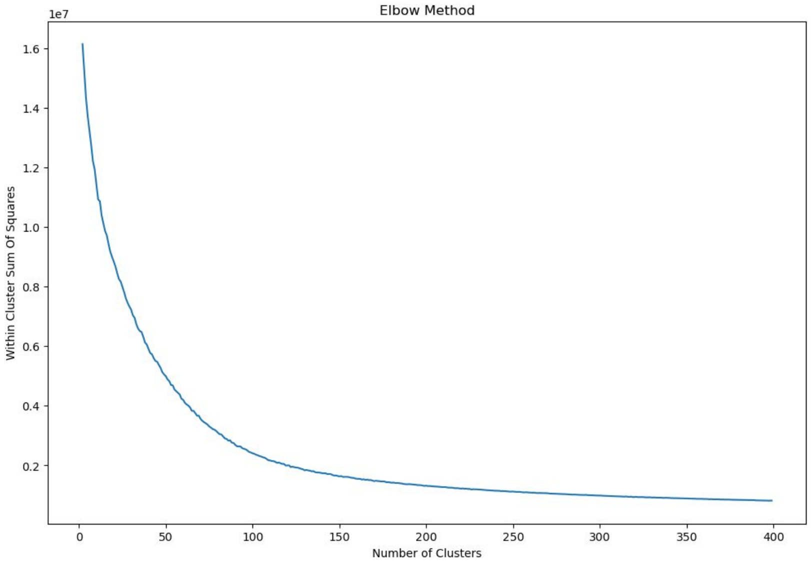
1. Skeletons -Clustering Visualization with t-sne

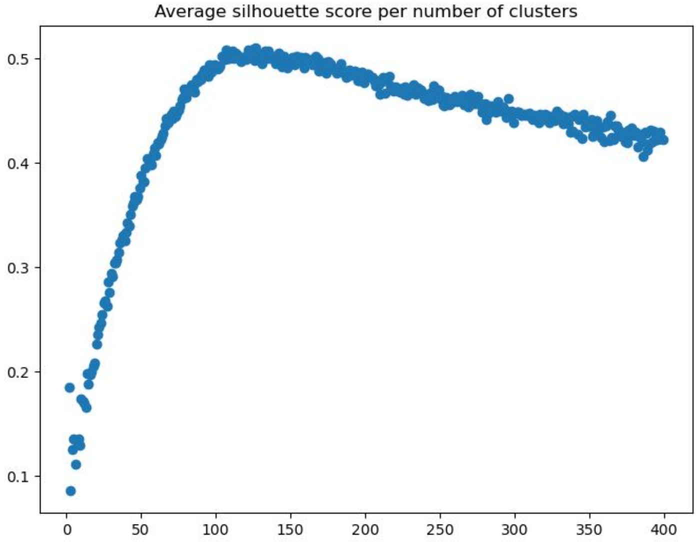


2. Images : Hierarchical Clustering – Ward Method



2. Images– Number of clusters





Clustering visualization only for embeddings data

Cluster

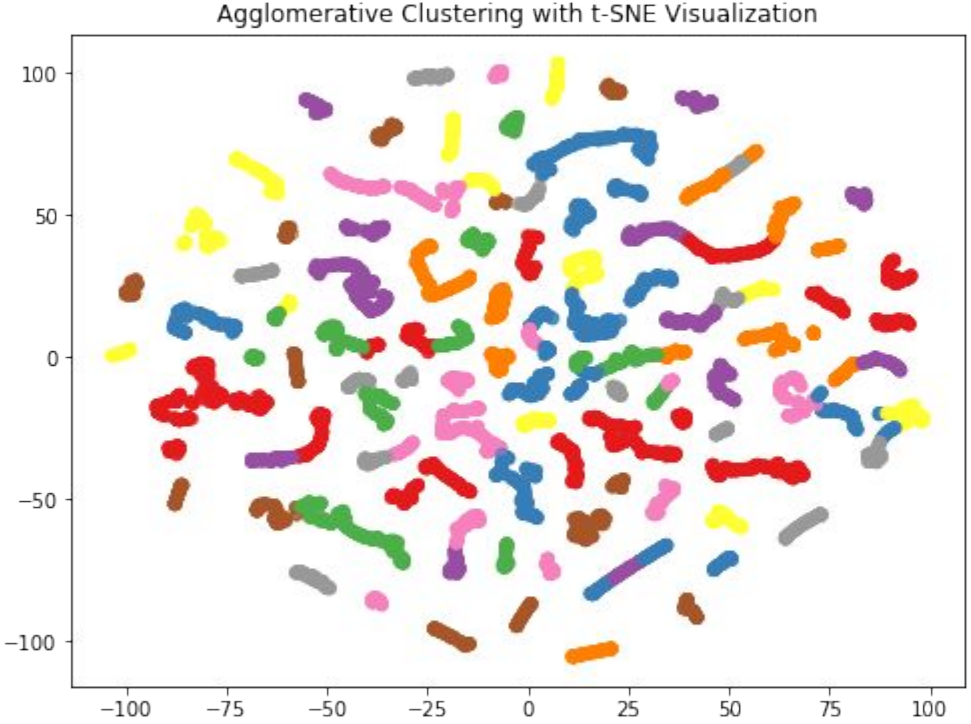
4

32

65

78

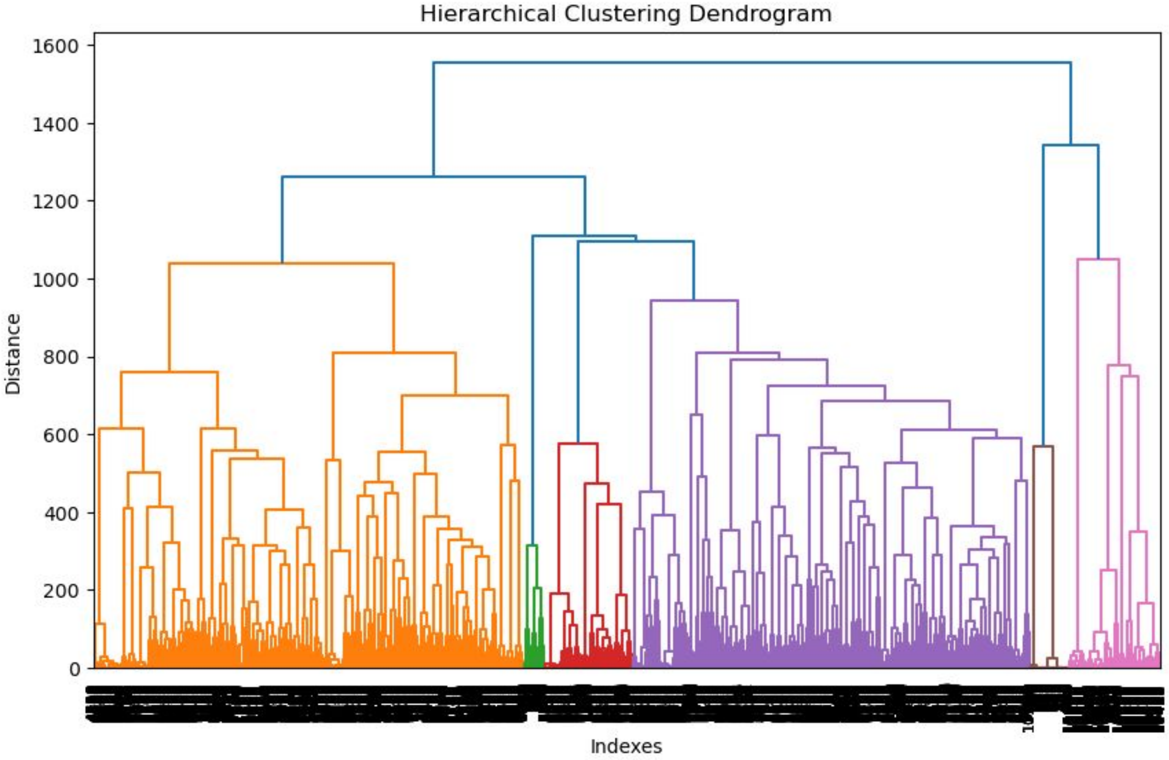
2. Images- Clustering Visualization with t-sne



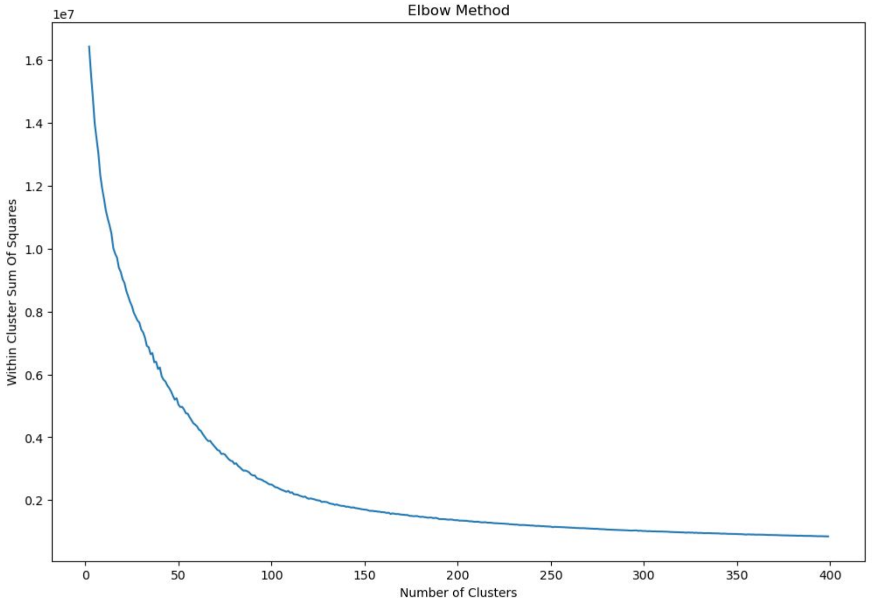
3. Skeletons + Images : Dataset Creation

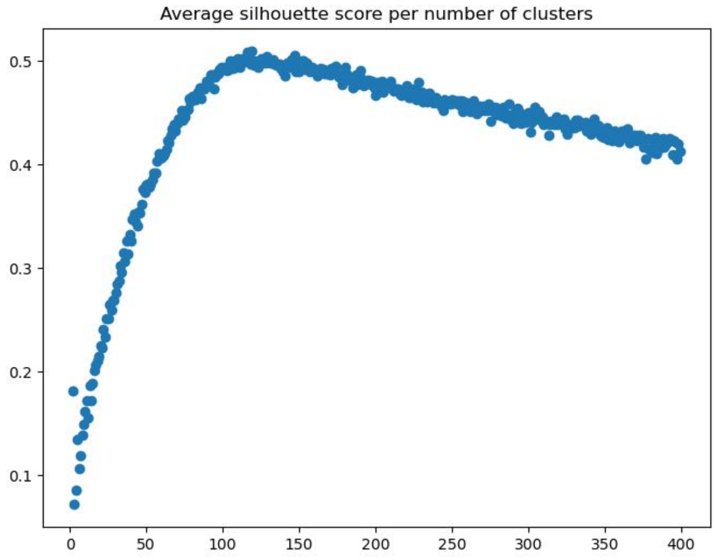
$$\begin{array}{rc} \text{Cluster} & \begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{matrix} \\ \mathit{Descriptor} = & \begin{bmatrix} 2 & 1 & 0 & 0 & 0 & 2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 2 \\ 2 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \end{array}, \qquad \mathit{Descriptor} \in \mathcal{R}^{n\times 8}$$

3. Skeletons + Images : Hierarchical Clustering – Ward Method



3. Skeletons + Images– Number of clusters





Clustering visualization for skeleton+embeddings data

Cluster

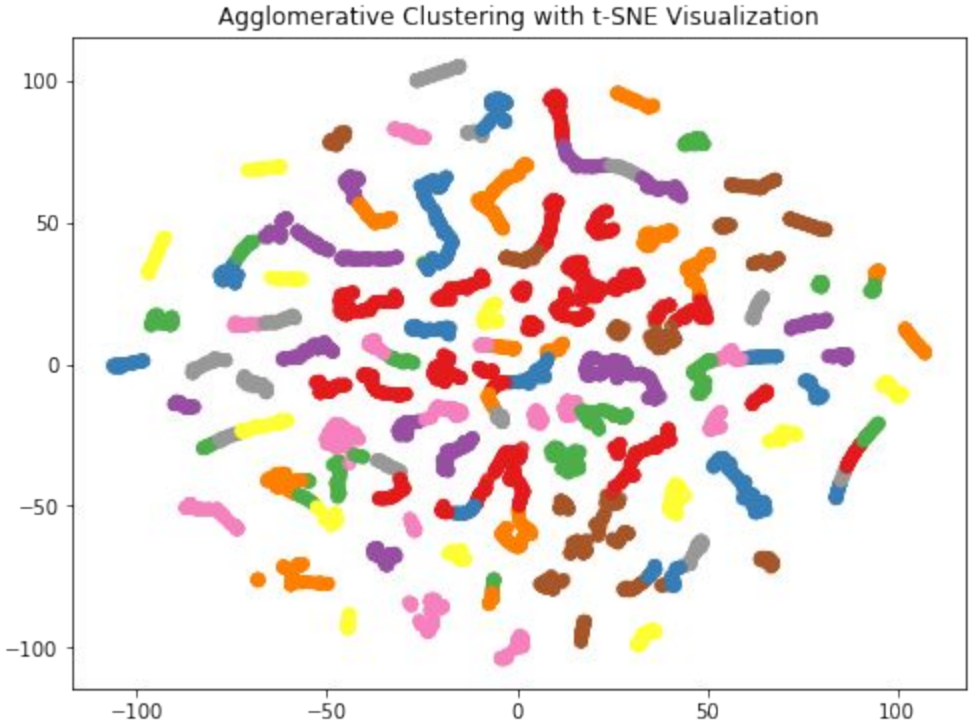
4

32

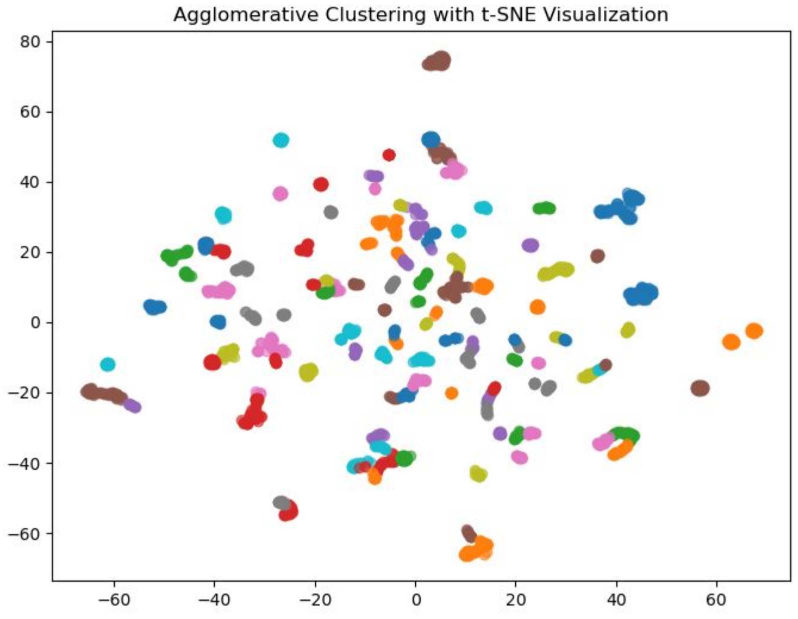
65

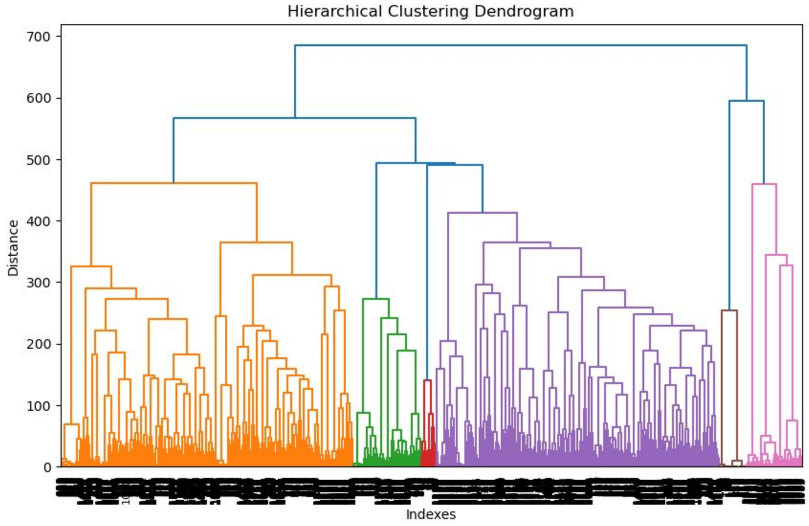
81

3. Skel + Images - Clustering Visualization with t-sne



Sampling every 5 frames





Sampling every 5 frames

Cluster

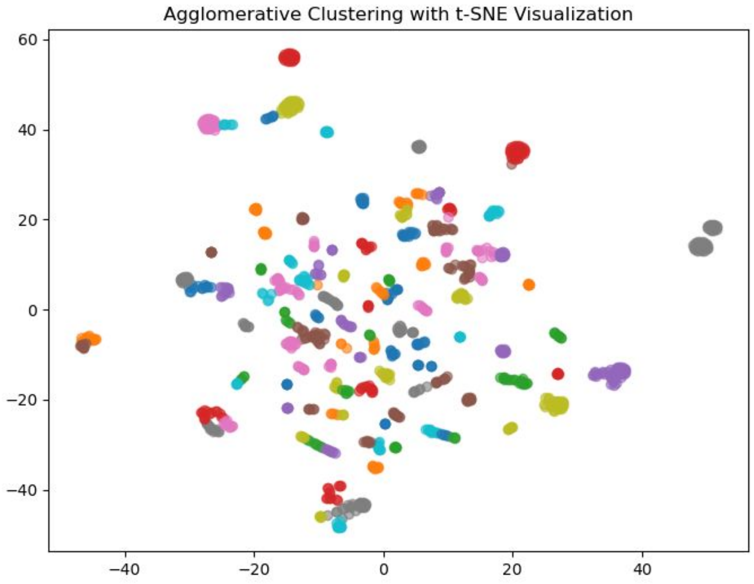
34

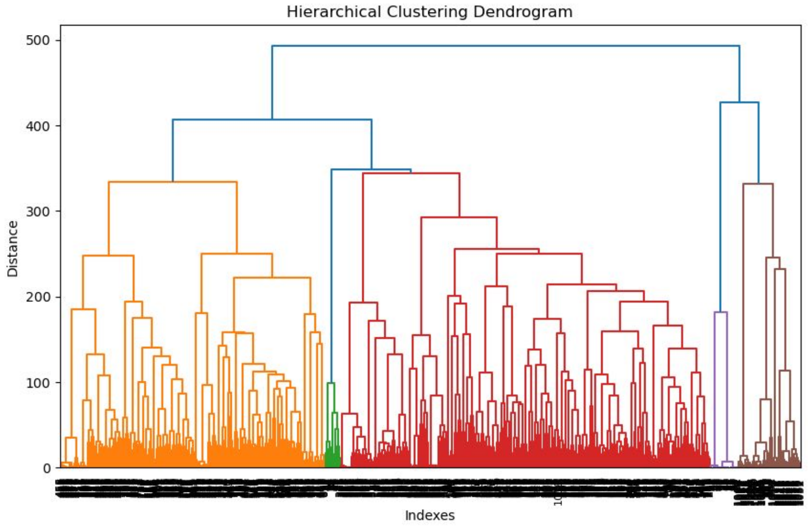
44

61

75

Sampling every 10 frames





Sampling every 10 frames

Cluster

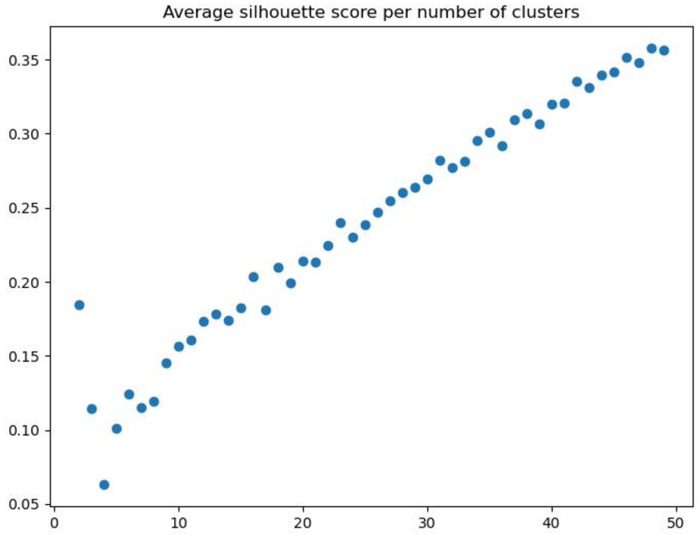
25

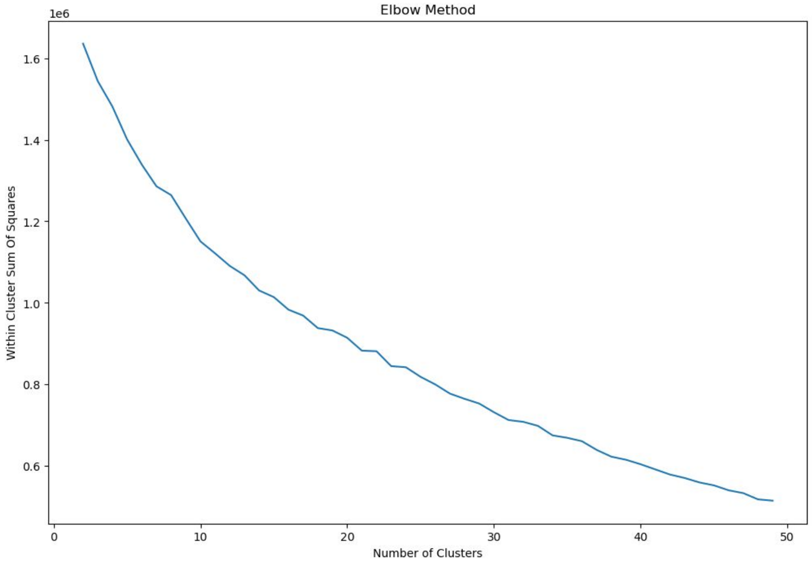
55

57

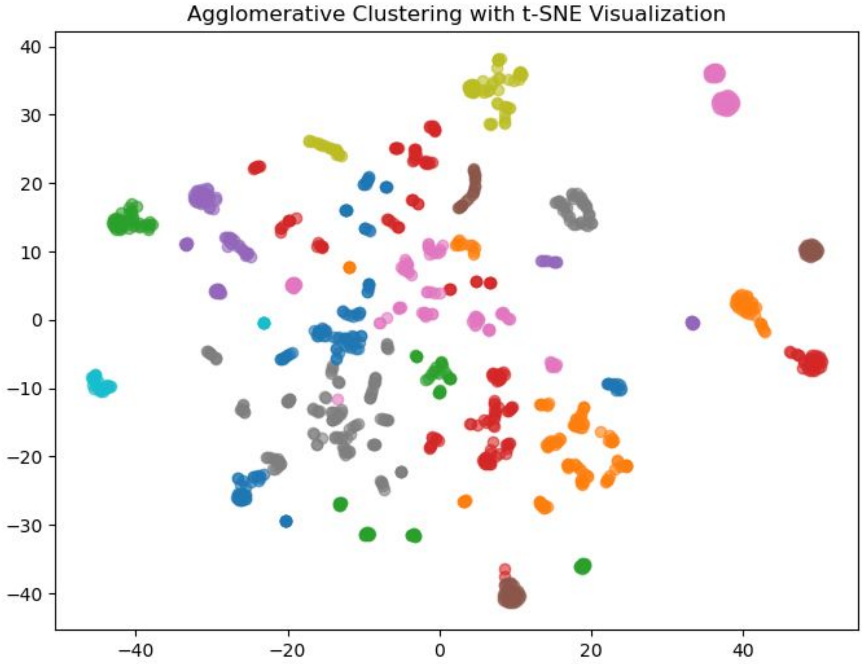
67

Sampling every 10 frames with fewer clusters





Sampling every 10 frames with 27 clusters



Sampling every 10 frames with 27 clusters

Cluster

9

10

16

17

Conclusions

- Complete pipeline with really satisfactory results for the 3 datasets considered
- There are several methods performed throughout the pipeline which will have a heavy impact on the results

Future work

- Classify frames based on skeleton data alone
- Test different methods for each pipeline module
- Add extra information to the skeleton and embeddings data
- Further reduce the image embeddings dimensionality

References

- Udell et al., *Generalized low rank models* 2016
- Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
- van der Maaten, Laurens & Hinton, Geoffrey. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research. 9. 2579-2605.
- Gomes, João Pedro. Lecture notes. Jun. 2023. url:https://fenix.tecnico.ulisboa.pt/disciplinas/PBDat/2022-2023/2-semestre/teoricas