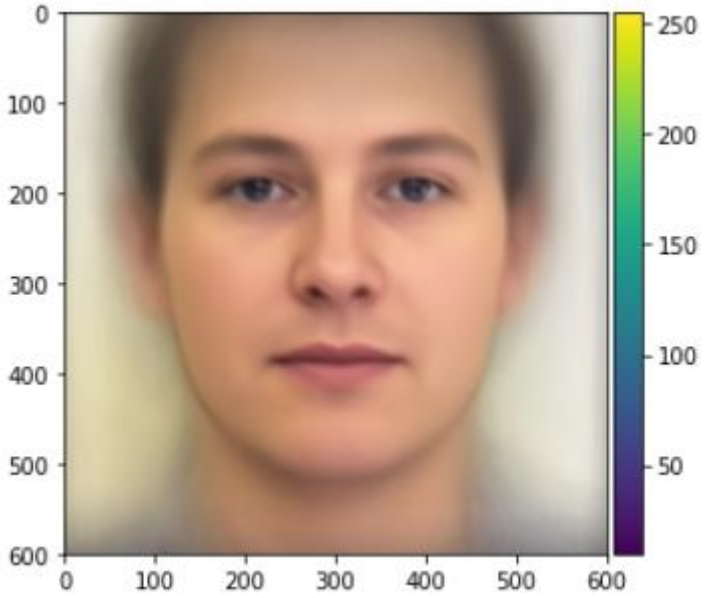
Machine Learning HW4 Report

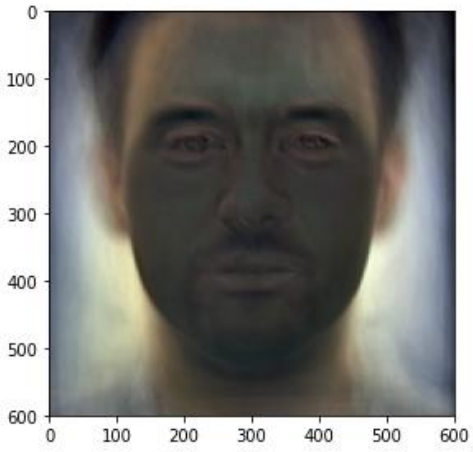
學號：B05611038 系籍：生機二 姓名：張育堂

PCA of colored faces

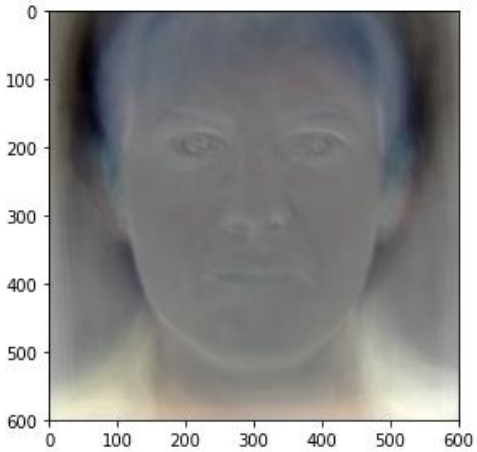
1. 請畫出所有臉的平均



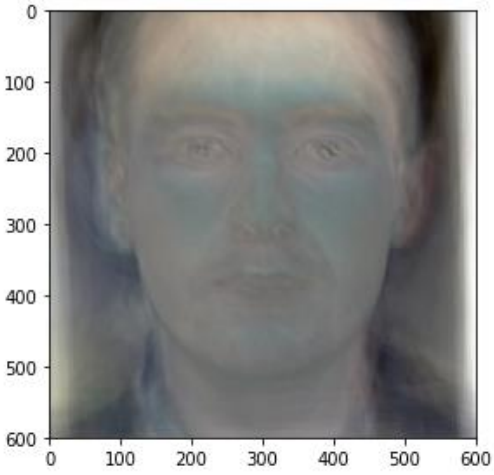
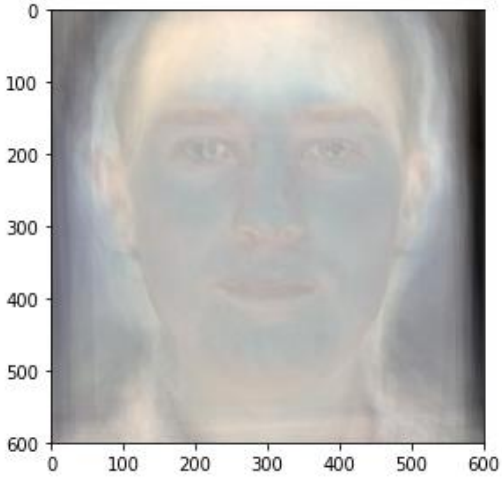
2. 請畫出前四個 Eigenfaces，也就是對應到前四大 Eigenvalues 的 Eigenvectors。



第一張



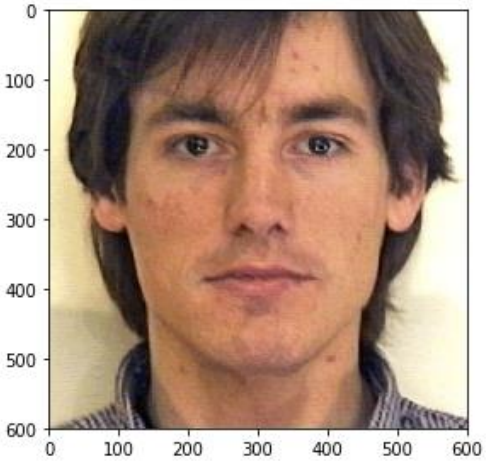
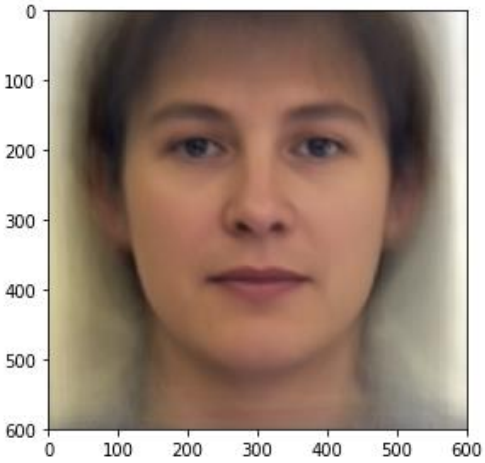
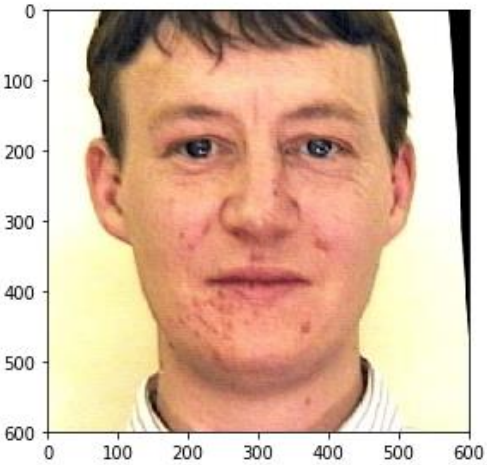
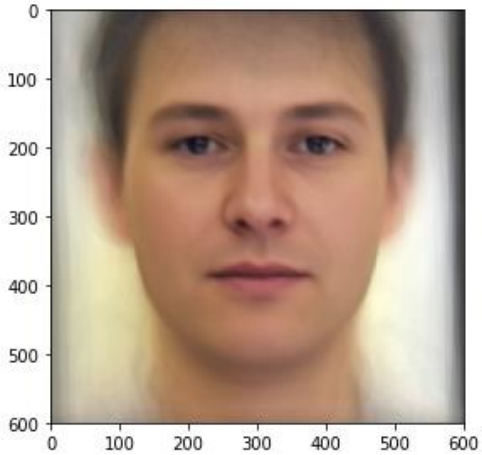
第二張

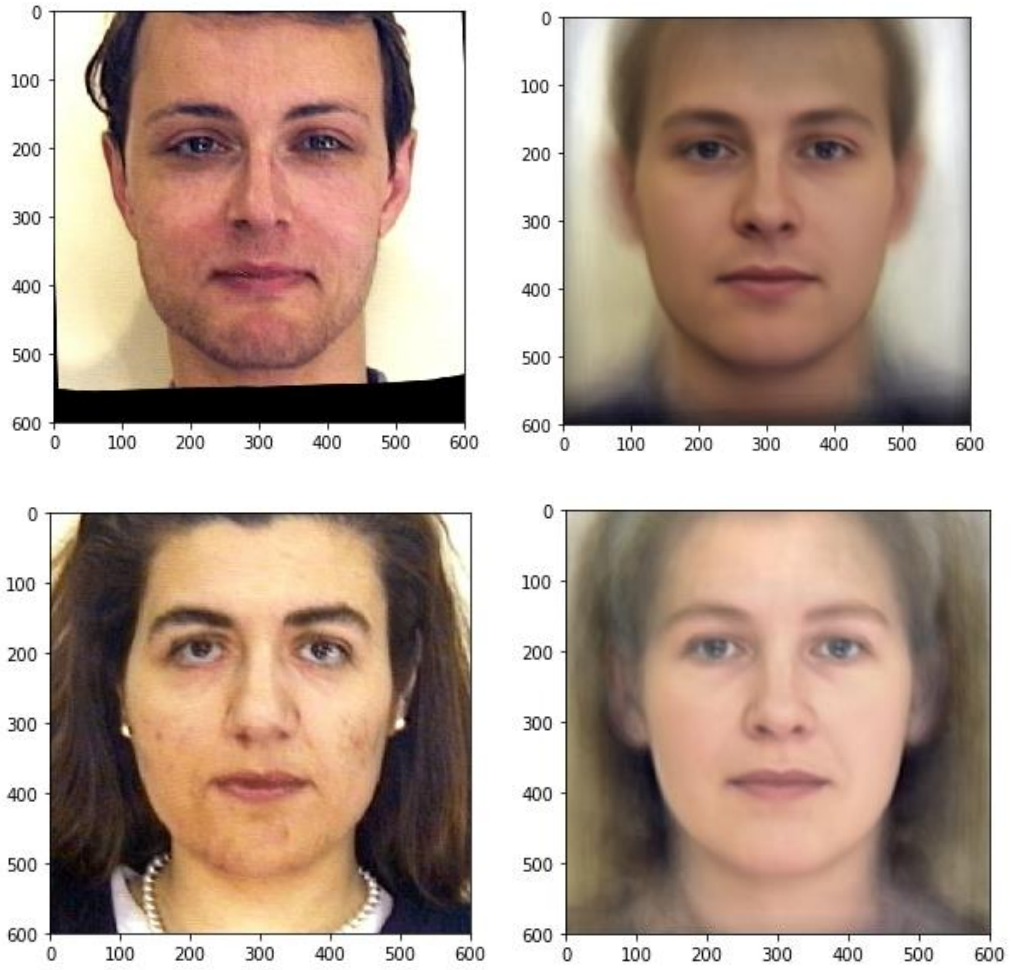
第三張　　第四張

3. 請從數據集中挑出任意四個圖片，並用前四大 Eigenfaces 進行 reconstruction，並畫出結果。

輸入　　重建結果

輸入　　重建結果



4. 請寫出前四大 Eigenfaces 各自所佔的比重，也就是 $\frac{S_i}{\sum S_j}$，請用百分比表示並四捨五入到小數點後一位。

第一高	第二高	第三高	第四高
4.1%	2.9%	2.4%	2.2%

Image clustering

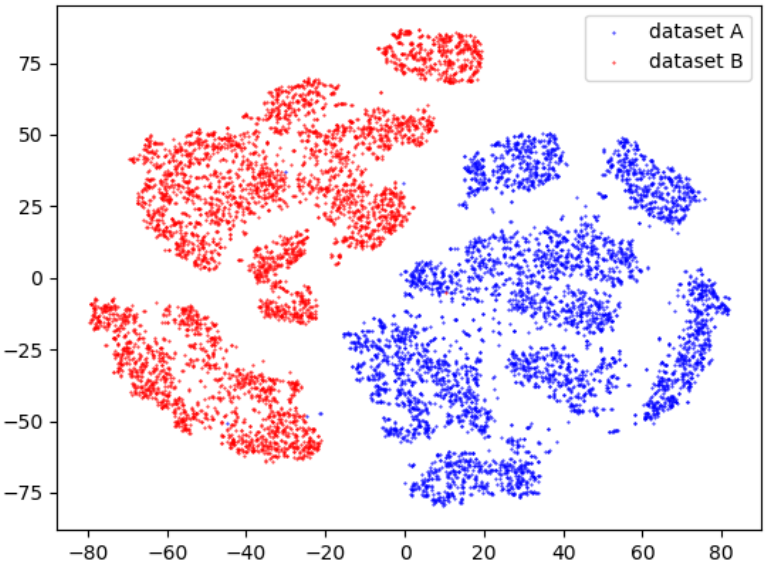
1. 請實作兩種不同的方法，並比較其結果。

Method	PCA+K-means	PCA+ t-SNE+ K-means
Score	1	0.719035

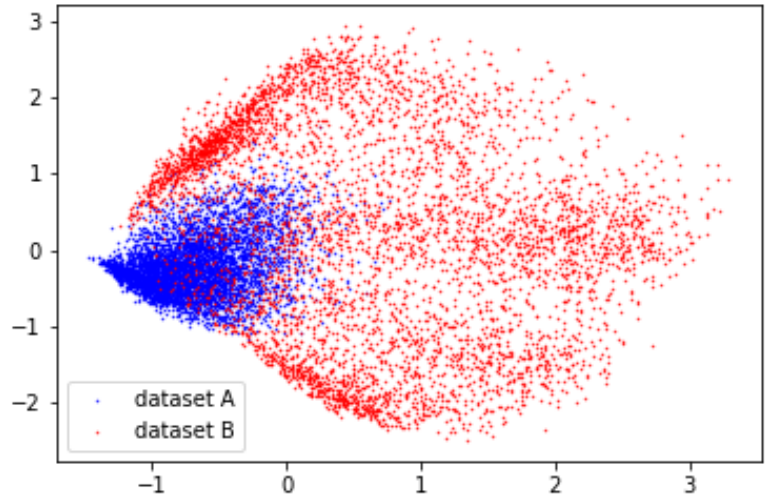
我的第一個方法是用 PCA 先降維到 300 以後，whiten 的參數有設置，之後直接用 k-means 作分類。第二個方法是使用 PCA 降維到 200 後(除了 n_component 以外全預設)，再用 t-SNE 降至 2 維，之後再讓 K-means 分類。

我想在第一個方法中有這麼好的結果，是因為 whiten 的參數功用有點相似於標準化，而在圖像的處理中，標準化的確是有助於分類的作法，這在上一份報告就有討論過了。

2. 預測 visualization.npy 中的 label，在二維平面上視覺化 label 的分佈



3. visualization.npy 中前 5000 個 images 來自 dataset A，後 5000 個 images 來自 dataset B。請根據這個資訊，在二維平面上視覺化 label 的分佈，接著比較和自己預測的 label 之間有何不同。



上圖是用 PCA(有設置 whiten)，直接將數據輸出在二維平面上的結果。而我在將其分類時是使用第一個 method，先將 visualization.npy 的數據使用 PCA(有設置 whiten)降到 30 維，用 K-means 標籤之後，在用其結果

跟前 5000 及後 5000 比數據進行比對，結果和第一小題一樣，預測和實際上沒有誤差，確認無誤差以後，我再用 t-SNE 降至 2 維並將圖輸出(第二題)。

由兩圖比較後可以看出，PCA + K-means 有將兩群明顯分開，雖然經過 t-SNE 輸出時有幾個不同群的點混雜在圖中，但因為實際確認無誤差後，只能認為是 t-SNE 視覺化雖然優秀但卻有可能無法真實呈現數據狀況。

Ensemble Learning

1. 請在 hw1/hw2/hw3 的 task 上擇一實作 ensemble learning，比較其與未使用 ensemble method 的模型在 public/private score 的表現並詳細說明你實作的方法。

我嘗試在 hw3 中實作 ensemble model。而我的 model 是使用四個類似 cnn16 架構的模型結構，先將四個 model 的 weight 取出來，並且再將 4 個 model 的 score 做平均之後再將其分類。

Model	Cnn16(ver1)	Cnn16(ver2)	Cnn16(ver3)	Cnn16(ver4)	Ensemble
Score	0.66912	0.666755	0.668285	0.664945	0.70632