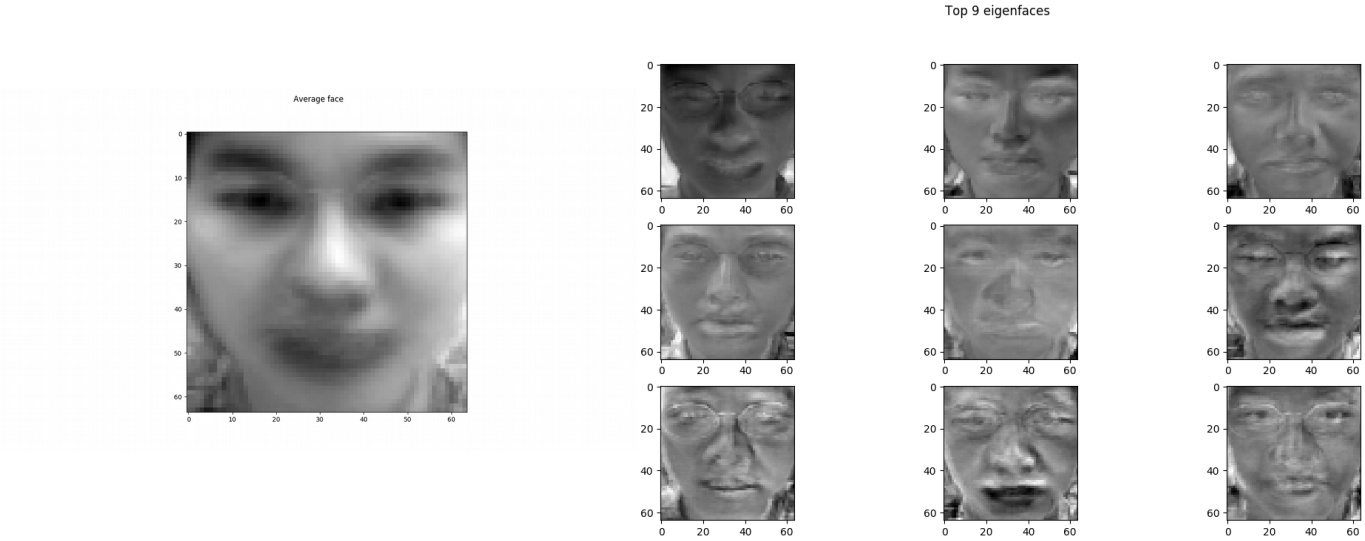
學號：B03705006 系級： 資管三 姓名：侯舜元

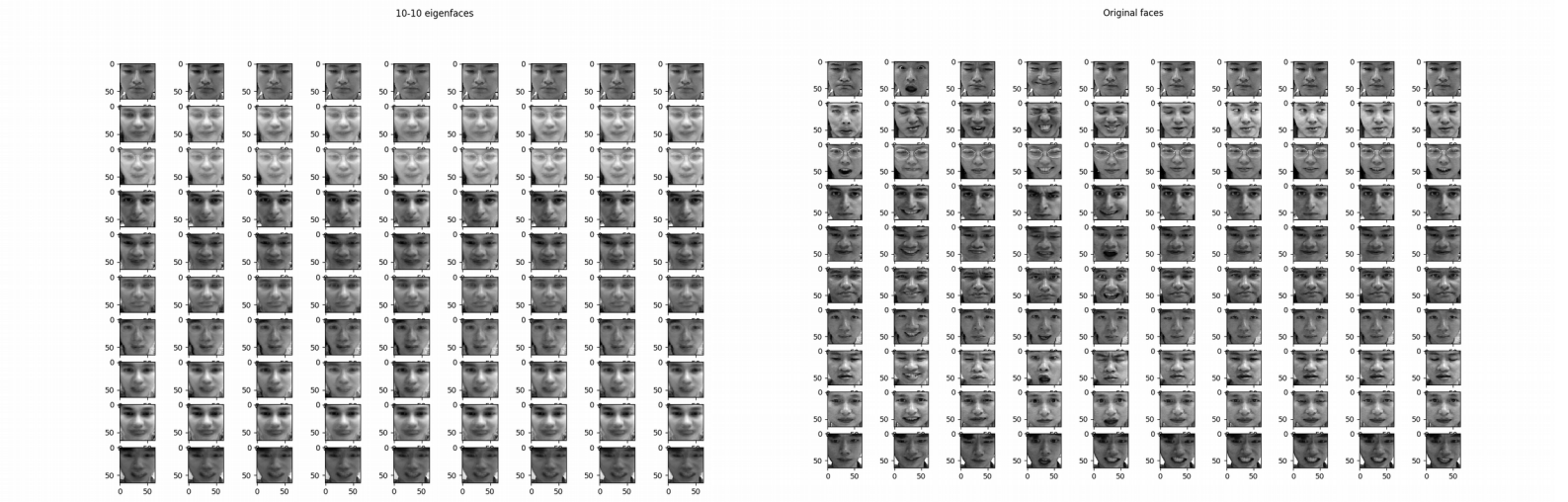
1.1. Dataset 中前 10 個人的前 10 張照片的平均臉和 PCA 得到的前 9 個 eigenfaces:

答： (左圖平均臉，右圖為 3x3 格狀 eigenfaces, 順序為 左到右再上到下)



1.2. Dataset 中前 10 個人的前 10 張照片的原始圖片和 reconstruct 圖 (用前 5 個 eigenfaces):

答：(左右各為 10x10 格狀的圖, 順序一樣是左到右再上到下)



1.3. Dataset 中前 10 個人的前 10 張照片投影到 top k eigenfaces 時就可以達到 < 1% 的 reconstruction error.
答: (回答 k 是多少)
(資料有做標準化/255)

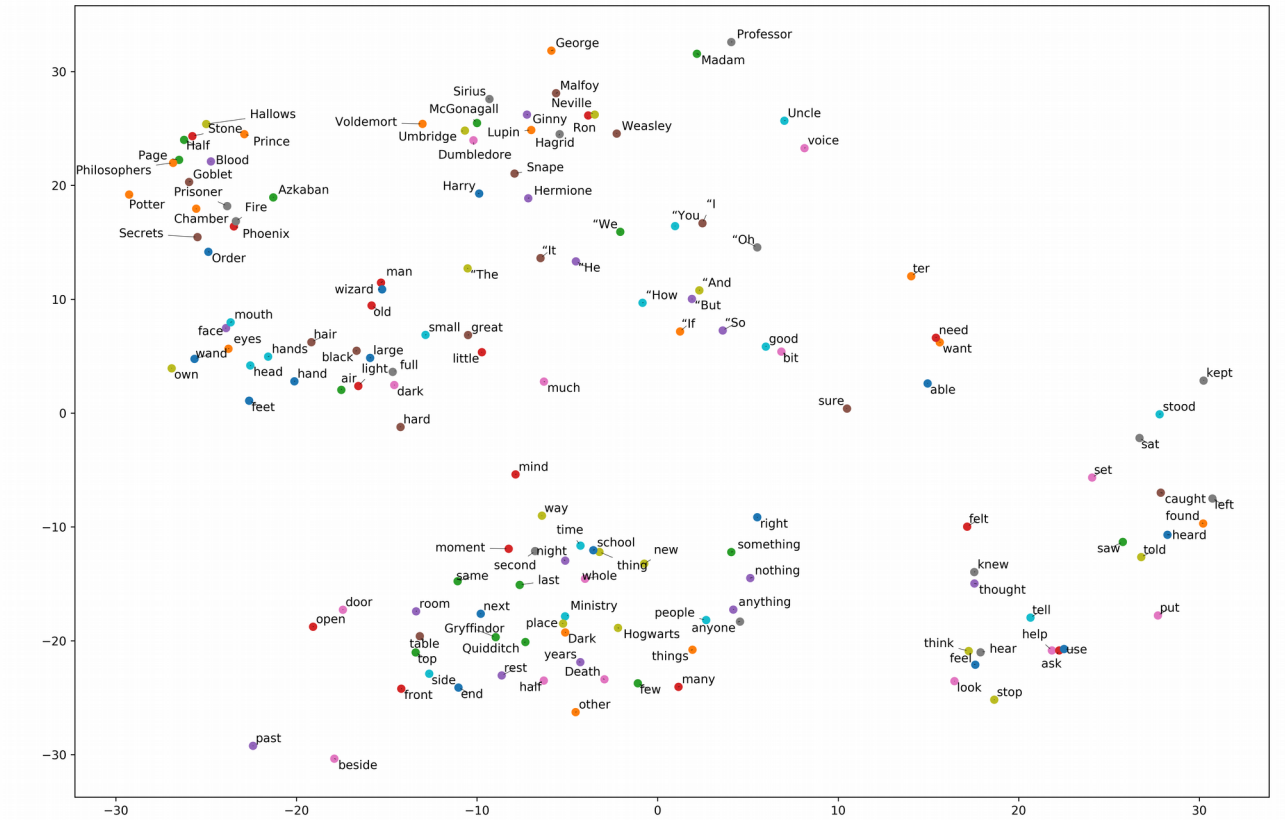
k	error
59	0.0103402779206
60	0.0100073230767
61	0.00971618411812
62	0.00941751220113

在 k = 61 時, 可以達到< 1% 的 reconstruction error.

2.1. 使用 word2vec toolkit 的各個參數的值與其意義:
答:

size = 100	word 的 vector 大小設為 100 維。有嘗試用更高維度表示，其投影圖就變得蠻雜亂不集中的
min_count = 5	至少要出現大於 5 次才會取樣
alpha = 0.05	Learning rate 設為 0.05(預設為 0.025)，稍稍調高讓分佈集中些
hs = 1	用 *Hierarchical Softmax*，可以使得原本要計算$\|V\|$次的訓練，縮減為 $log2\|V\|$次，提升了訓練速度
threads = 1	用 1 個 thread 去執行
window = 5	字跟字之間的 max skip length 設為 5(預設)
verbose = True	Training 時輸出結果

2.2. 將 word2vec 的結果投影到 2 維的圖:
答: (圖)



2.3. 從上題視覺化的圖中觀察到了什麼?
答:
可以看到圖上分成幾個 cluster。左上方字的群集，都跟哈利波特系列的大主題有關。例如：鳳凰會、阿茲卡班、魔法石、混血王子等，都是書名或是幾個重要的主題、元素。而正上方的 *cluster*，主要是人名的群集。有趣的是，可以看到同一家族的人名的距離比較小，群聚在一起。左中的 *cluster* 是跟人相關的簡單名詞集合。正下方的 *cluster* 詞類比較雜，推測是跟 *Hogwarts*, *Gryffindor* 等學校生活中常出現的字彙的群集。右下方的 cluser 則是各種動詞和動作。

另外幾點比較有趣的觀察:
(1) 明顯同義字的距離會非常近、近乎重疊，例如說(need,want), (knew, thought),
(2) 連續相關連動作、或是同類型詞彙距離也會很近，(open, door), (Professor, Madam), (face, mouth, hand, wand) 魔杖可能也是巫師身體的一部分
(3) 魔法部淪陷後”dark’’跟”Ministry”變得非常接近
(4) 主角三人組並沒有想像中接近，中間夾雜了石內卜等人

3.1. 請詳加解釋你估計原始維度的原理、合理性，這方法的通用性如何?
答:
由於使用 K-D tree 不是個好方法，在做 NearestNeighbors 時我們使用 ball tree。而我們選擇先用題目給的條件 h is a dimension sampled from [60, 79] uniformly 去 generate 一些原生維度為 1~60 的資料，再拿該資料去 train linear SVR。最後利用 SVR 去找 eigenvalues 跟原始維度的回歸模型。將 SVR(C=10) c 的參數微調到 10。 最後在 predict 時，將整數+-0.2 之間的維度 round。

原理跟合理性:
原先題目 oracle network 的方法可能跟 SVD 有些類似，在維度增加時可能有相似的 NearestNeighbors。故找出 eigenvalues 後再用 SVR 去做推測。

通用性:
受亂數影響大，每次執行的結果可能誤差甚大。

3.2. 將你的方法做在 hand rotation sequence datatset 上得到什麼結果? 合理嗎? 請討論之。
答: