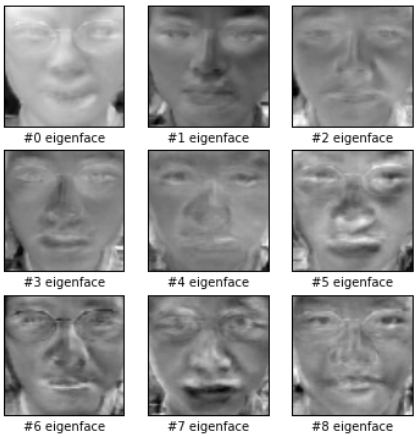
Machine Learning HW4

學號：R04922169 系級：資工所碩二 姓名：楊智偉

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)

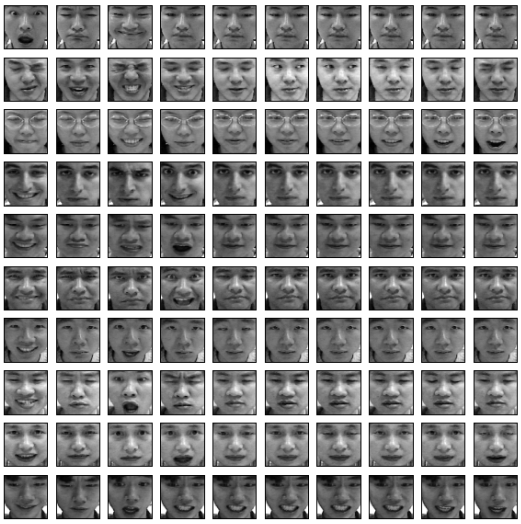
Average face



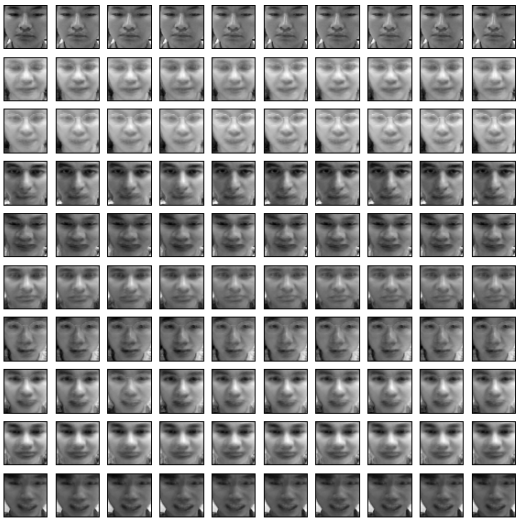
Eigenfaces

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)



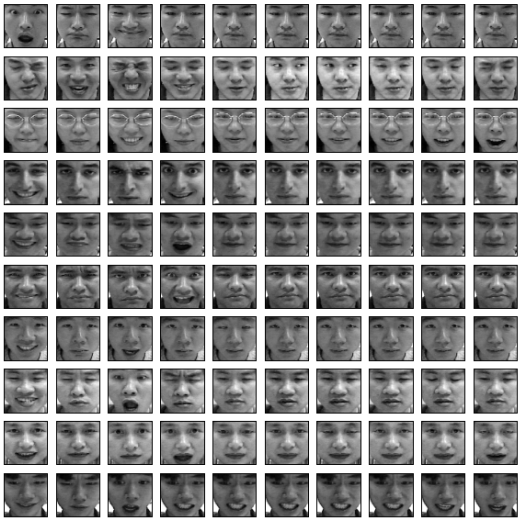
Original images



Reconstructed images

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：經過計算後，我發現 k 達到 59 之後可以達到小於 1%的 reconstruction error。



Original images

Reconstructed images with k=59

2.1 使用 word2vec toolkit 的各個參數的值與其意義：

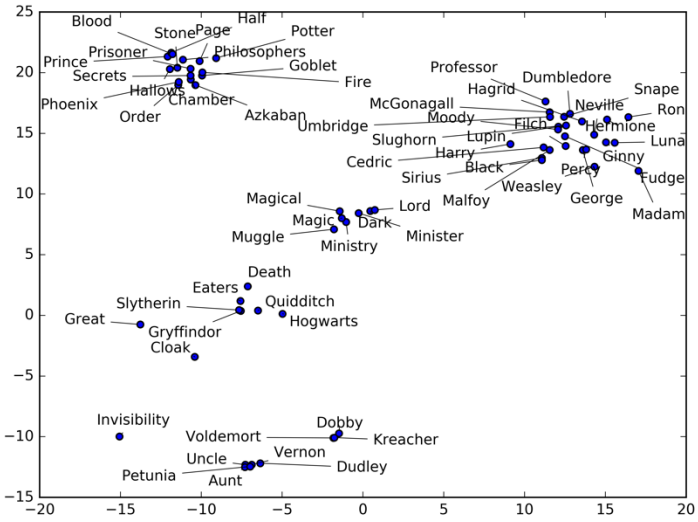
答：

- train：training data 的檔案。
- output：存放 training 結果的的檔案。
- cbow：是否使用 cbow。0 表示使用 skip-gram 模型，1 表示使用 cbow 模型。
- size：表示輸出的詞特徵 vector 的維度。
- min_count：可以對字典做截斷，頻率少於 min_count 次數的單詞會被丟棄掉。
- window：training 時的窗口大小，代表當前詞會前後各考慮幾個詞。
- negative：>0 的時候用 negative sampling，代表會使用多少的 noise words。
- iter_：training 的迭代次數。
- alpha：learning rate，學習速率。

  我所使用到的參數如下圖左所列。

```
MIN_COUNT = 10
WORDVEC_DIM = 500
WINDOW = 5
NEGATIVE_SAMPLES = 5
ITERATIONS = 10
MODEL = 0
LEARNING_RATE = 0.01
```

word2vec parameters



word2vec result

2.2 將 word2vec 的結果投影到 2 維的圖：
答：如上圖右。

2.3 從上題視覺化的圖中觀察到了什麼？
答：

由上題的 2 維圖，我們確實可以看到某些相關的詞在空間中的位置相當的近。先觀察最左上的那堆，'Blood', 'Prisoner', 'Prince', 'Secret', 'Stone', 'Chamber' 等等的詞彙都是 Harry Potter 七集書名的關鍵字。

另外，在看到最下面的那一小群，Harry 小時候是住在 'Aunt', 'Uncle' 的家裏，Aunt 和 Uncle 的名字分別是 'Petunia' 'Dudley', 'Vernon' 'Dudley'，所以這幾個詞是比較相關的。而最後右上最大群的部分，我認為都是與主角們較熟識人物名稱，諸如 'Moody', 'Dumbledore', 'Weasley', 'George', 'Ginny', 'Luna', 'Snap', 'McGonagall', 'Ron' 等等重要的角色。

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：

$$\mathbb{R}^{d_i} \xrightarrow{\text{ELU}} \mathbb{R}^{h_i} \xrightarrow{\text{ELU}} \mathbb{R}^{100} \xrightarrow{\text{Linear}} \mathbb{R}^{100}$$

根據我們所知道的資訊，data points 是由 normal distribution 產生，且 W, b 的數值也是使用 normal distribution 隨機產生。所以嘗試大膽的假設：若原始的資料的維度越低，那麼經過 transform 之後所得到的 100 維資料的 variance 也會越小。

而我們還知道 di 的範圍是[1, 60]，所以我就將此 200 個 datasets 先計算 variance，再透過 k means 分為 60 個 clusters，由 center 小至大一一配上 1~60 的 dimension。估算原始維度的時候，再將 data 比對 60 個 clusters 進行內插即可。

此方法的缺陷就是通用性很低，因為是利用了這些 dataset 是用 normal distribution 來產生，又知道實際上他們是 1~60 維度的資料。對於現實世界的資料而言就沒那麼合適了。

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

我將 dataset 轉換為 10*10 的圖片，以 100 維對應原本的演算法，並且估計出 dataset 的維度是 28。很顯然的這並不合理，因為此 dataset 的圖片很單純簡單，不應該這麼高維。如前所述，此方法的通用性太低，對於不同性質的 dataset 會有非常大的影響，而且真實的 data 就不是使用 normal distribution 產生的規律 dataset 了。