

Tarea 2

Viridiana Itzel Méndez Vásquez

21 de abril de 2025

1. EJERCICIO 3

Considera el conjunto de datos Labelled Faces in the Wild (LFW) que consiste en fotografías de rostros recolectados de internet y contenido en `sklearn`. Algunos rostros identificados, tienen varuas fotos incluídas en el dataset. Vamos a considerar solo aquellas personas que tienen al menos 70 fotografías de su rostro, también, vamos a considerar el tamaño original de la imagen (125×94).

```
1 from sklearn.datasets import fetch_lfw_people
2 lfw_people = fetch_lfw_people(min_faces_per_person=70, resize=1)
```

Esto resulta en 1288 imágenes que pertenecen a alguna de las etiquetas:

```
1 for name in lfw_people.target_names:
2     print(name)
3
4 Ariel Sharon
5 Colin Powell
6 Donald Rumsfeld
7 George W Bush
8 Gerhard Schroeder
9 Hugo Chavez
10 Tony Blair
```

1. Separa un conjunto de entrenamiento (80 %) y prueba (puedes usar la función `train_test_split` de `sklearn.model_selection`), por ejemplo:

```

1 from sklearn.model_selection import train_test_split
2 X_train, X_test, y_train, y_test, names_train, names_test = train_test_split(X, y, target_names[y], test_
size=0.2, random_state=42)

```

donde antes, tuviste que declarar `X`, `y`, `target_names` (ve la documentación de `fecfh_lfw_people`). Obtén las eigenfaces del conjunto de entrenamiento. Visualiza los scores de los primeros dos componentes principales ¿Encuentras patrones interesantes?

2. Proyecta los datos de prueba en los componentes principales. Verifica si se “ubican” en su “individuo” correspondiente al graficarlos en los primeros dos componentes principales.
3. Usa el método del vecino más cercano para identificar a un “sujeto” de prueba en las imágenes de entrenamiento. Usa la distancia euclídea en el espacio de los p componentes principales. Decide qué valor de p usar. El objetivo es obtener algo como lo que se muestra en la Figura 3 del archivo de la Tarea:
¿Puedes identificar correctamente a los sujetos usando éste criterio? ¿Qué tanto influye el valor de p ?
4. Considera una(s) imágen(es) que no están en la base de datos ¿qué se te ocurre para prevenir casos como los que muestran en la Figura 4 del archivo de la Tarea?

1.1. SOLUCIÓN:

El conjunto de datos Labelled Faces in the Wild (LFW) con el que estaremos trabajando, es un conjunto de fotografías de rostros de distintas personas, dicho conjunto se encuentra contenido en `sklearn`. Se estará considerando además, personas que tengan al menos 70 fotografías en la base de datos. Cada fotografía está identificada por una etiqueta y a través de esta etiqueta, por su nombre, el tamaño de estas fotografías es de 125×94 pixeles. En la Figura 1.1 se muestran las primeras cuatro personas con más de 70 fotografías en el dataset teniendo un total de 1288.

Al realizar exploración en los datos, notamos que están desbalanceados pues contamos con 530 fotografías de George W Bush, contra 236 fotografías de Colin Powell, menos de la mitad del número de veces en el que aparece el rostro de George W Bush, mientras que Hugo Chavez es la persona con menos fotografías en el dataset, con un total de 71. Este desbalance en el número de fotografías genera problemas al momento de realizar predicciones, además de no darnos mucha información.

A) APPLICACIÓN DE PCA

A continuación se separaron los datos en un conjunto de entrenamiento con el 80 % de los datos y un conjunto de prueba con el resto, de forma aleatoria, de manera que



Figura 1.1: Ejemplos de rostros de personas contenidos en el dataset Labelled Faces in the Wild

nuestro conjunto de entrenamiento cuenta con 1030 fotografías, mientras que el conjunto de prueba cuenta con 258 fotografías. Antes de aplicar PCA estandaricé las imágenes, aunque en esta ocasión no es necesario debido a que todas las imágenes cuentan con las mismas dimensiones.

El método de PCA se aplicó al conjunto de entrenamiento para 100 componentes, en la Figura 1.2 podemos ver la proporción de la varianza explicada con respecto al número de componentes, observemos que si queremos el 80 % necesitamos los primeros 42 componentes principales, pero si con el 70 % es suficiente, necesitamos los primeros 18 componentes principales, esto nos será de gran utilidad más adelante.

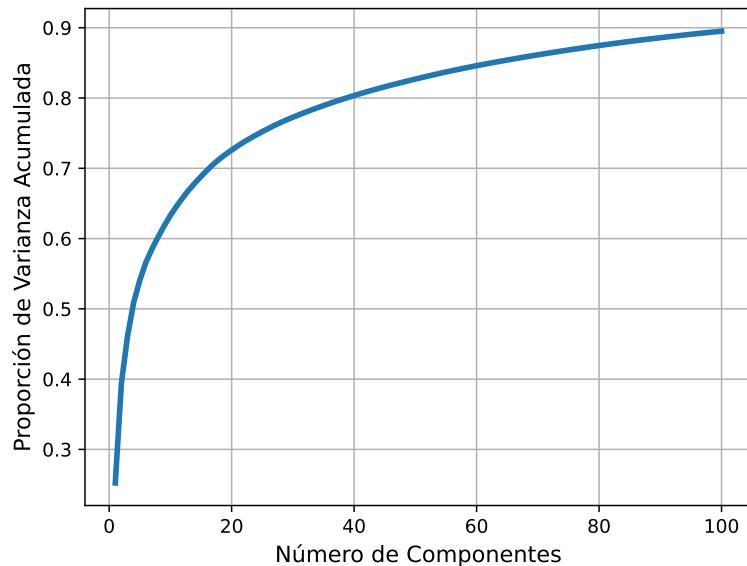


Figura 1.2: Varianza explicada al aplicar PCA al dataset de las fotografías para 100 componentes

Posteriormente obtuvimos las primeras 12 eigenfaces, las cuales se muestran en la Figura 1.3. Observemos que cada eigenface extrae información relevante de la cara, pero po-

demos notar que las primeras cuatro eigenfaces no tienen bien definidos rasgos faciales, estas primeras eigenfaces capturan los cambios de iluminación en el rostro, y a partir del Eigenface 4 se extraen los rasgos característicos del rostro.

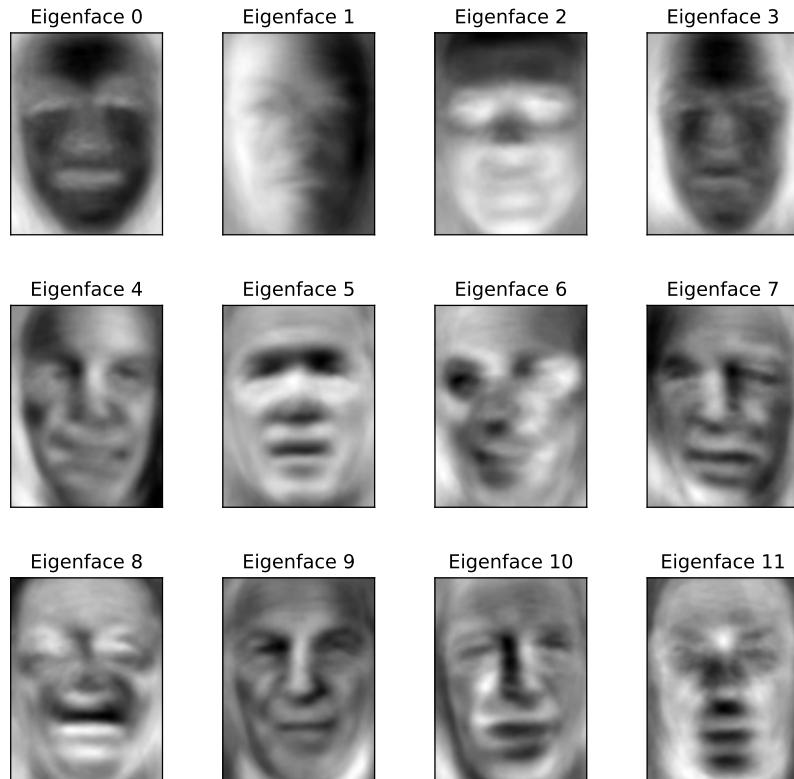


Figura 1.3: Primeras 12 eigenfaces del conjunto de entrenamiento.

Luego, se graficaron los scores de los primeros dos componentes principales, para poder identificar patrones se colocaron las fotografías originales en los puntos correspondientes, dicha gráfica se muestra en la Figura 1.4. Podemos notar que si nos movemos hacia la izquierda, la mayoría de los rostros miran hacia ese lado, además que sonríen o tienen expresión más alegre, mientras que si nos movemos hacia el lado derecho, la mayoría de los rostros también miran hacia ese lado, pero tienen expresiones más serias, el ceño fruncido, los cuales son los patrones que se pueden notar.

b) PROYECCIÓN DE LOS DATOS DE PRUEBA

Para proyectar los datos de prueba en los componentes principales, transformamos los datos de prueba y los graficamos de forma conjunta con los datos de entrenamiento transformados, todo esto en los dos primeros componentes principales. Se construyó una gráfica interactiva diferenciando por el tipo de marcador entre la proyección de los datos de entrenamiento y de prueba, además de colorear por nombre, pero principalmente, de

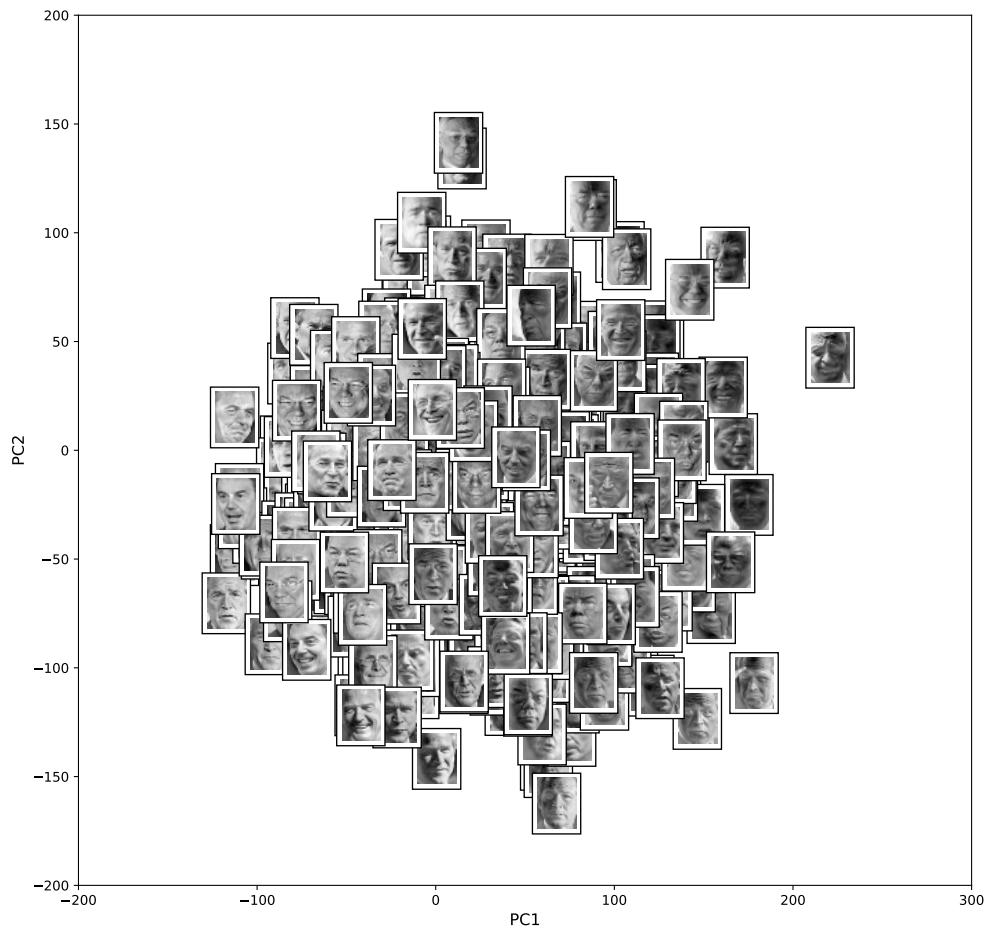


Figura 1.4: Gráfica de los scores de los dos primeros componentes principales.

forma que al posicionarse en un punto se pueda desplegar el nombre de la persona. En la Figura 1.5 se muestra sólo la gráfica diferenciando entre las proyecciones. Aquí se presenta un problema con lo mencionado en la introducción al ejercicio sobre el desbalance de los datos, hay muchas fotografías de George W Bush, en la gráfica el color morado lo representa, por lo que lo podemos ver por toda la gráfica, esto no nos da mucha información sobre si los datos de prueba se ubican realmente en su individuo al tener tantas repeticiones de la misma persona, por ejemplo, Hugo Chavez es la persona con el menor número de fotografías en los datos, en la gráfica lo podemos identificar con el color azul marino, sólo en un punto de prueba se encuentra con puntos de entrenamiento al rededor que corresponden a él, sin embargo en el resto de puntos no se identifica a la misma persona al rededor, pues también en el conjunto de entrenamientos se tienen pocos datos correspondientes a él.

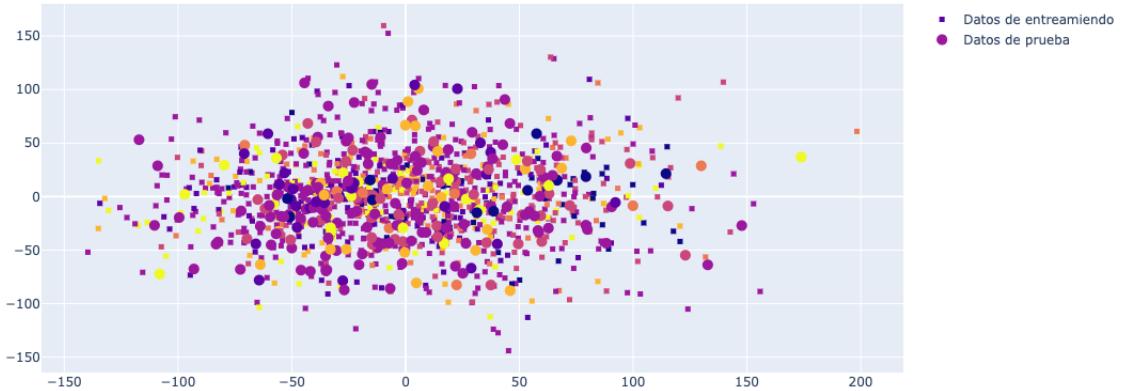


Figura 1.5: Proyección de los datos de prueba en los dos primeros componentes principales.

c) MÉTODO DEL VECINO MÁS CERCANO PARA IDENTIFICAR A UN “SUJETO” DE PRUEBA EN LAS IMÁGENES DE ENTRENAMIENTO.

Para este caso, se tomó a alguna persona de prueba y se proyectó en el espacio de p componentes, considerando diferentes valores de p , después se midió las distancias con respecto a los datos de entrenamiento en el mismo espacio para encontrar así al más cercano, la distancia utilizada fue la distancia euclídea. Se observan casos interesantes, en la Figura 1.6a se muestra la identificación de una fotografía de Tony Blair, para $p = 42$, recordemos que con ese número de componentes se recupera un 80 % de la varianza total y en este caso, su vecino más cercano es efectivamente el mismo individuo con expresiones faciales diferentes, por otro lado, cuando se considera $p = 100$ en el que se recupera aproximadamente el 90 % de la varianza total y por tanto, en la proyección se pueden notar los rasgos más definidos, como se observa en la Figura 1.6c, en este caso su vecino más cercano es Donald Rumsfeld, pero pongamos atención en la expresión que tiene en la fotografía, la posición del rostro es hacia la misma dirección, aunque la mirada de Donald Rumsfeld es en dirección distinta a la del sujeto de prueba, además ambos tienen la boca abierta, su expresión en general es muy parecida, en este caso sus proyecciones son muy similares, por eso su identificación, finalmente para $p = 400$ tenemos más del 90 % de la varianza total, mostrado en la Figura 1.6e, nos dice que su vecino más cercano es George W Bush, aunque el sujeto es equivocado, las expresiones faciales son más parecidas aún, sin embargo ya es estamos considerando un número considerable de componentes, en comparación con los 42 que utilizamos primero y que sí nos devolvió al sujeto correcto. Ahora pongamos atención en la Figura ?? y 1.6d, donde se quiere identificar a Colin Powell, para este sujeto a pesar de tomar los componentes principales que nos devuelvan

el 80% y 90% de la varianza total, en ninguno de los dos casos el vecino más cercano nos devuelve a la persona correcta, sin embargo la expresión facial de los vecinos más cercano es cada vez más parecida a la de nuestro sujeto de prueba.

Los sujetos que se consideraron anteriormente aunque no cuentan con el número suficiente de fotografías como el caso de George W Bush, tampoco se encuentran dentro de los individuos con el menor número de fotografías.

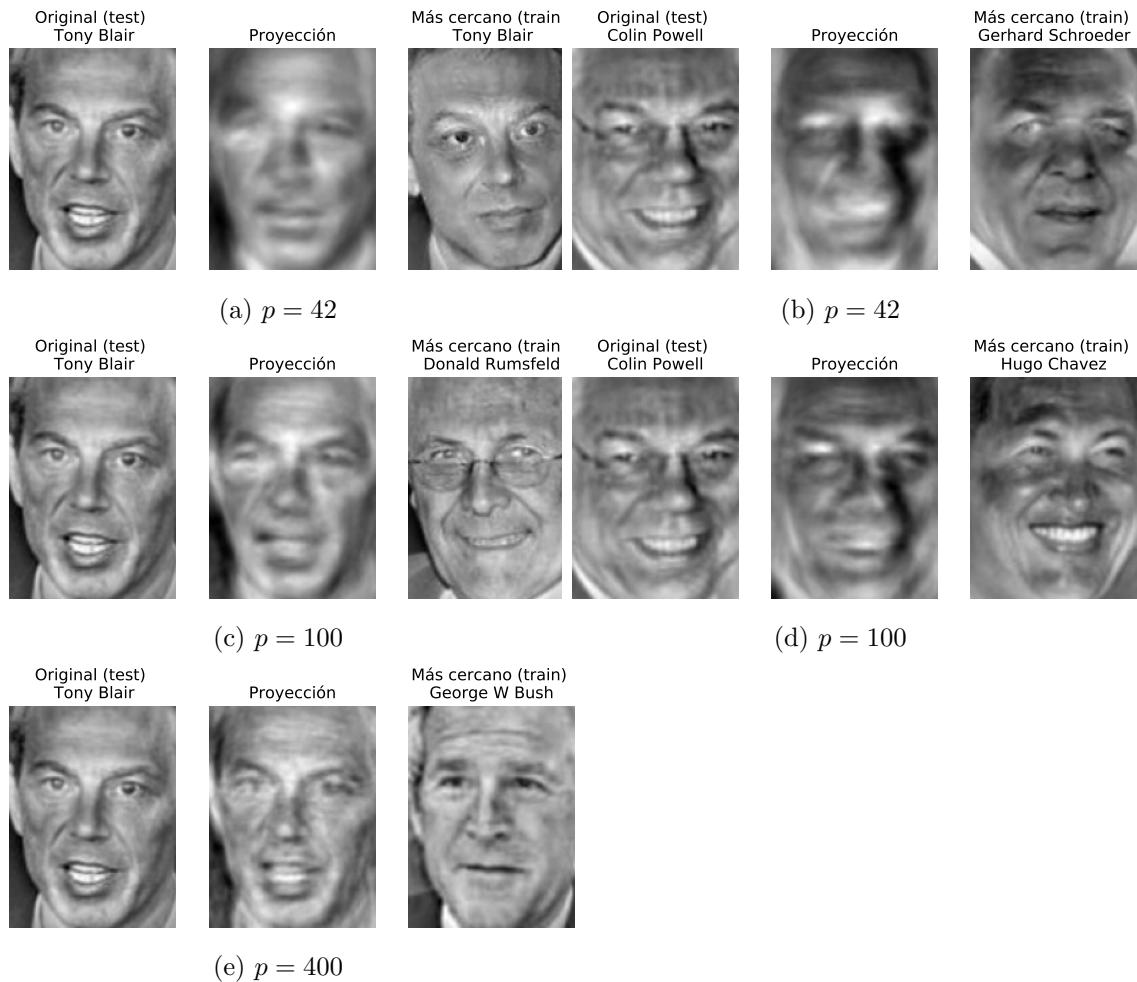


Figura 1.6: Identificación de un individuo de prueba usando el vecino más cercano en el espacio de los primeros p componentes principales, para diferentes valores de p .

Otros casos interesantes se muestran en la Figura 1.7, donde 1.7a y 1.7c corresponden a la identificación de George W Bush, pero con diferentes fotografías, recordemos que es la persona con el mayor número de fotografías en la base de datos, se puede notar que para el mismo valor de p , en el primer fotografía, su vecino más cercano es efectivamente el mismo George Bush, sin embargo para la segunda fotografía no sucede lo mismo,

pero su vecino más cercano tiene una expresión muy parecida a la del sujeto. En las Figuras 1.7b, 1.7d queremos identificar a Ariel Sharon, quien es la segunda persona con el menor número de fotografías en la base de datos lo que puede hacer difícil su correcta identificación, para los valores de $p = 18, 42, 100$, la proyección evidentemente cambia, pero el vecino más cercano en todos los casos es el mismo, que es lo que se quiere presentar en 1.7b. Por último, cuando el valor de $p = 400$ el vecino más cercano es George Bush. Este último es un caso en el que no se va a ubicar en su individuo, lo que se debe en mayor parte al desbalanceo de los datos.

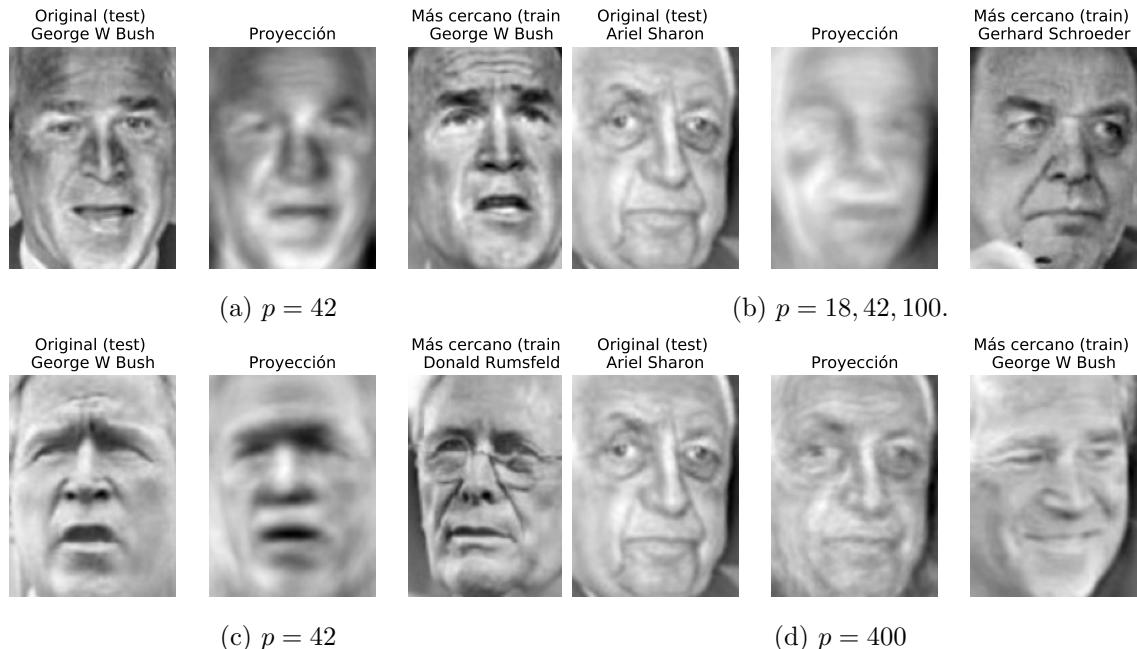


Figura 1.7: Identificación de un individuo de prueba usando el vecino más cercano en el espacio de los primeros p componentes principales, para diferentes valores de p .

La identificación de un individuo de prueba en el espacio de los p componentes es muy interesante porque se puede decir que está identificando bien a los individuos, en el sentido en el que tienen una gran cantidad de rasgos en común, expresiones, posición, etc., esto debido a que las proyecciones son muy similares y por eso los está identificando aunque no sea la persona correcta. El valor de p influye en qué tanto parentezco van a tener y esto es claro, pues entre más componentes consideremos, se espera que el porcentaje de la varianza total extraída sea cada vez mayor.

d) IDENTIFICACIÓN DE UN INDIVIDUO FUERA DE LA BASE DE DATOS

Para este caso, en la base de datos LFW se consideraron personas con al menos 40 fotografías para tener a personas que no se encuentran en los datos de entrenamiento, se busca en dicha base de datos para que tengan las mismas dimensiones que el resto.

Consideramos a Serena Williams, Arnold Schwarzenegger y a Jennifer Capriati, puesto que estandrizamos los datos de entrenamiento, estandarizamos también los datos correspondientes a las personas mencionadas. En la Figura 1.8 se muestra la persona a identificar, su proyección y su vecino más cercano.

Notemos que en todos los casos la proyección no es clara, no se puede identificar nada, sin embargo se coloca en algún punto del espacio de los p componentes principales. En el caso de $p = 18$, para Arnold Schwarzenegger, (ver Figura 1.8a) el vecino más cercano que devuelve es Donald Rumsfeld, mientras que para Serena Williams y Jennifer Capriati (ver Figuras 1.8c, 1.8e) devuelven exactamente a la misma persona con la misma fotografía. Algo similar sucede en el caso en el que se utilizan 42 componentes, la proyección es indistinguible y la persona que devuelve el vecino más cercano es exactamente la misma para todos los casos.

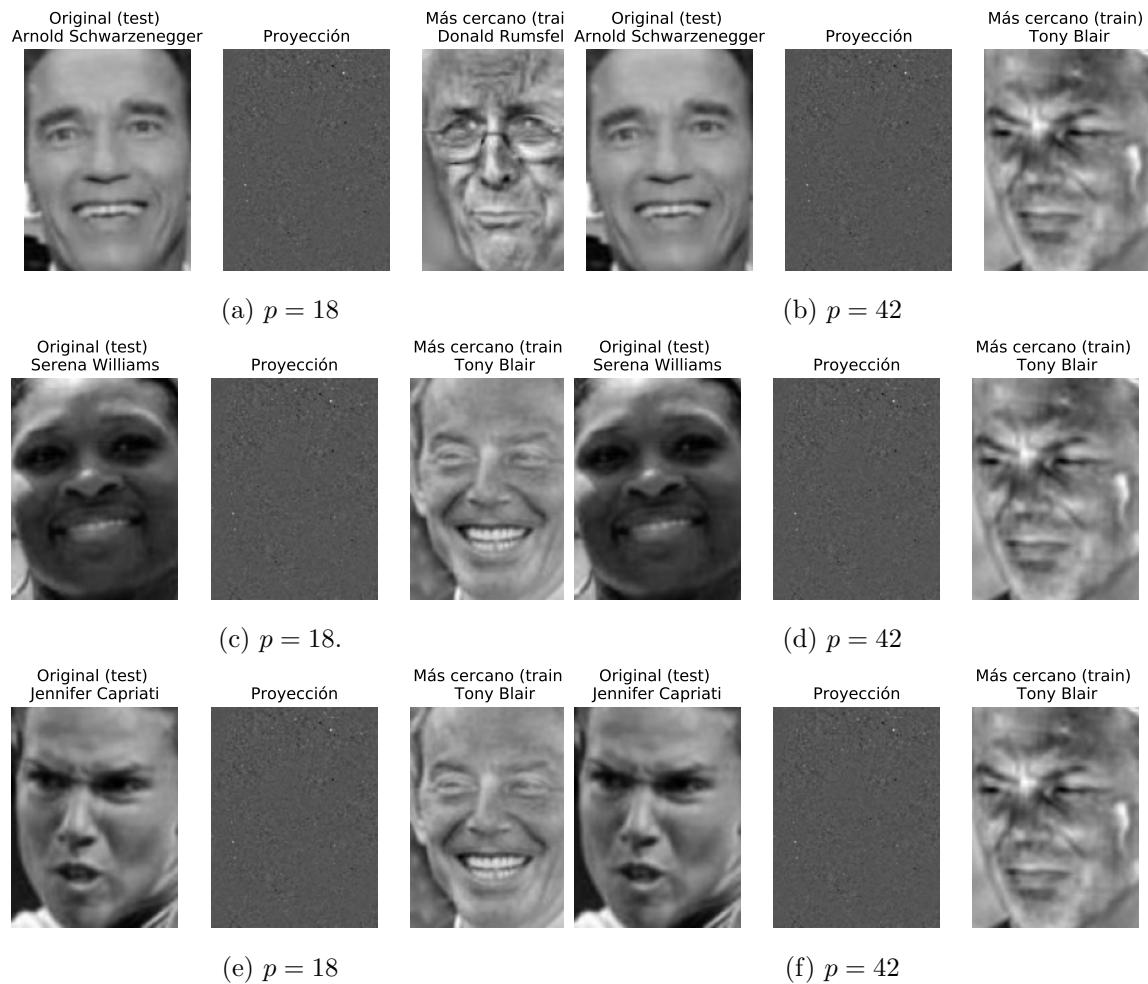


Figura 1.8: Identificación de un individuo de prueba externo a la base de datos usando el vecino más cercano en el espacio de los primeros p componentes principales, para diferentes valores de p .

2. EJERCICIO 4

Dado un conjunto de datos centrados $\mathbf{X}_{n \times d}$ vimos que hacer PCA, es realizar la descomposición espectral de la matriz de covarianzas muestral, que puede estimarse como $\mathbf{S} = \mathbf{X}'\mathbf{X}$ (omitimos el coeficiente $n - 1$).

Ahora, considera la matriz $K_{n \times n} = \mathbf{XX}'$.

1. Muestra que es equivalente realizar PCA en \mathbf{S} o en \mathbf{K} , es decir, que $(\lambda^{-1/2}\mathbf{X}\mathbf{u}, \lambda)$ es un par eigenvector-eigenvalor normalizado de \mathbf{K} , y a su vez, $(\lambda^{-1/2}\mathbf{X}^T\mathbf{v}, \lambda)$ es un par eigenvector-eigenvalor normalizado de \mathbf{S} , donde \mathbf{u} y \mathbf{v} son vectores propios de \mathbf{S} y \mathbf{K} , respectivamente.
2. Verifica experimentalmente el resultado del inciso previo en el conjunto de imágenes LFW que usaste en el ejercicio anterior, ¿En qué casos es recomendable usar \mathbf{K} ?

2.1. SOLUCIÓN:

a) EQUIVALENCIA EN REALIZAR PCA EN \mathbf{S} O EN \mathbf{K} .

Demostración. Sean \mathbf{u} y \mathbf{v} vectores propios de \mathbf{S} y \mathbf{K} , respectivamente, donde $\mathbf{S} = \mathbf{X}'\mathbf{X}$ y $K_{n \times n} = \mathbf{XX}'$. Dado que \mathbf{u} es vector propio de \mathbf{S} correspondiente al valor propio, digamos, λ . Por definición de eigenvalor y eigenvector, se satisface que

$$\mathbf{Su} = \lambda u, \quad \text{lo que implica que} \quad \mathbf{X}'\mathbf{X}\mathbf{u} = \lambda \mathbf{u} \quad (2.1)$$

Multiplicando \mathbf{X} por la derecha en la igualdad anterior, se obtiene que $\mathbf{XX}'\mathbf{X}\mathbf{u} = \lambda \mathbf{X}\mathbf{u}$, de donde, asociando, se verifica que

$$\mathbf{K}(\mathbf{X} \mathbf{u}) = \lambda \mathbf{X}\mathbf{u}$$

es decir, $\mathbf{X}\mathbf{u}$ es un eigenvector de \mathbf{K} correspondiente al eigenvalor λ . Resta obtener su norma, para esto veamos lo siguiente

$$||\mathbf{X}\mathbf{u}||^2 = (\mathbf{X}\mathbf{u})'(\mathbf{X}\mathbf{u}) = \mathbf{u}'\mathbf{X}'\mathbf{X}\mathbf{u}$$

al sustituir 2.1, se obtiene que

$$||\mathbf{X}\mathbf{u}||^2 = \mathbf{u}'\lambda\mathbf{u} = \lambda\mathbf{u}'\mathbf{u} = \lambda.$$

Pues el eigenvector \mathbf{u} se considera ya normalizado. Con todo lo anterior se prueba que $(\lambda^{-1/2}\mathbf{X}\mathbf{u}, \lambda)$ es un par eigenvector-eigenvalor normalizado de \mathbf{K} .

Por otro lado, considerando \mathbf{v} vector propio normalizado de \mathbf{K} , se garantiza, por definición que $\mathbf{K}\mathbf{v} = \lambda\mathbf{v}$, y al sustituir la forma de K , se tiene $\mathbf{XX}'\mathbf{v} = \lambda\mathbf{v}$, al multiplicar la expresión anterior por la derecha por la matriz \mathbf{X}' , y asociando, se obtiene

$$\mathbf{X}'\mathbf{X}\mathbf{X}'\mathbf{v} = \lambda\mathbf{X}'\mathbf{v} = \mathbf{S}\mathbf{X}'\mathbf{v} = \lambda\mathbf{X}'\mathbf{v},$$

es decir, $\mathbf{X}'\mathbf{v}$ es un eigenvector de \mathbf{S} correspondiente al eigenvalor λ . Encontremos su norma como sigue

$$\|\mathbf{X}'\mathbf{v}\|^2 = (\mathbf{X}'\mathbf{v})'(\mathbf{X}'\mathbf{v}) = \mathbf{v}'(\mathbf{X}\mathbf{X}'\mathbf{v}) = \lambda\mathbf{v}'\mathbf{v} = \lambda.$$

Por lo tanto, $(\lambda^{-1/2}\mathbf{X}^T\mathbf{v}, \lambda)$ es un par eigenvector-eigenvalor normalizado de \mathbf{S} . \square

b) COMPROBACIÓN EXPERIMENTAL EN EL CONJUNTO DE DATOS LFW

Para la realización de este ejercicio se utilizó la misma base de datos Labelled Faces in the Wild, se estandarizaron los datos y se procedió a separar un conjunto de entrenamiento, como en el ejercicio anterior.

Se sabe que si se realiza PCA sobre el conjunto de entrenamiento, se estará haciendo uso de la matriz \mathbf{S} , de modo que para utilizar la matriz $\mathbf{K}_{n \times n}$ se aplica PCA al conjunto de datos transpuesto, posteriormente, para verificar que efectivamente es equivalente usar PCA en \mathbf{S} o en \mathbf{K} , a los componentes obtenidos de utilizar K , se transforman a componentes de \mathbf{S} mediante $\lambda^{-1/2}\mathbf{X}^T\mathbf{v}$ donde \mathbf{v} son los componentes principales de \mathbf{K} , posteriormente se obtienen las eigenfaces generadas al aplicar PCA al conjunto de entrenamiento sin transponer, las cuales se muestran en la Figura 2.1, mientras que en la Figura 2.2 se muestran las eigenfaces que se obtienen de los componentes principales haciendo uso de \mathbf{K} al aplicar la transformación. El objetivo de presentar dichas eigenfaces es observar que son las mismas, es decir, que los componentes principales obtenidos al transformar, proporcionan la misma información que los componentes principales obtenidos de la matriz sin transponer.

Notemos que los pares de Eigenfaces 0 – 7, 1 – 0, 2 – 1, 3 – 2, 4 – 3, 5 – 4, 6 – 5, 7 – 6 corresponden a las mismas, donde el primer número señala a las eigenfaces en la Figura 2.1 y el segundo a las que se encuentran en la Figura 2.2, pero se presta especial atención en los pares 0 – 7, 3 – 2, 4 – 3 y 7 – 6 donde el contraste es diferente de las eigenfaces es diferentes, sin embargo representan la misma información, sólo es cuestión de la dirección de los componentes principales.

¿En qué casos es recomendable usar \mathbf{K} ?

Se observa que al tener $X_{n \times d}$, donde se tienen n observaciones y d variables, conviene utilizar \mathbf{K} en los casos en los que se tienen un número menor observaciones que de variables, porque así se trabaja con una matriz de dimensión menor.

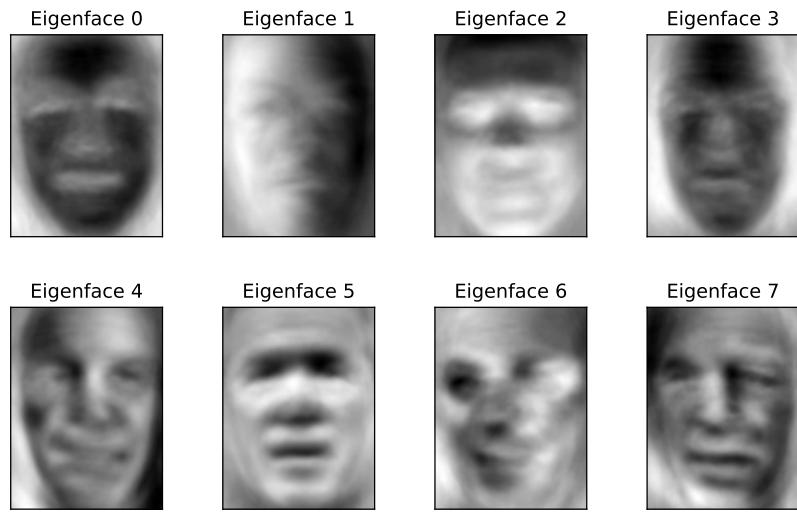


Figura 2.1: Primeras 12 eigenfaces del conjunto de entrenamiento haciendo uso de la matriz \mathbf{S} .

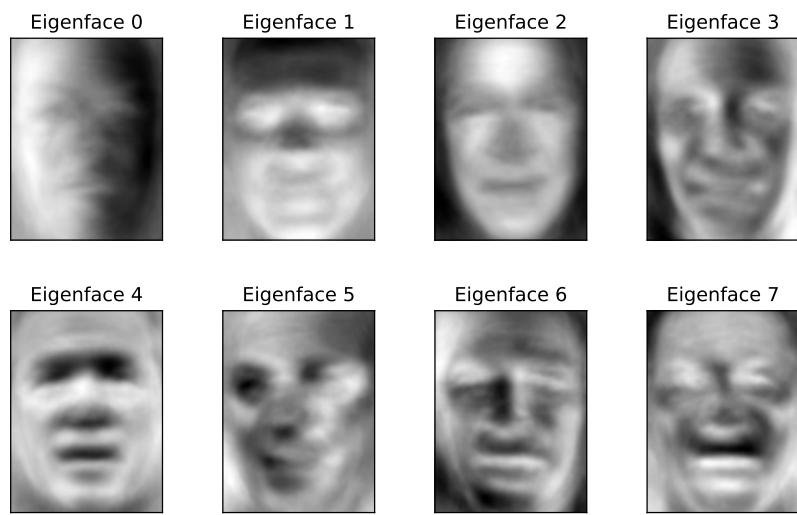


Figura 2.2: Primeras 12 eigenfaces del conjunto de entrenamiento haciendo uso de la matriz \mathbf{K} .