Ingeniare. Revista Chilena de Ingeniería
ISSN: 0718-3291
facing@uta.cl
Universidad de Tarapacá
Chile

Aguilar-Torres, Eduardo; Bekios-Calfa, Juan
Clasificación de género utilizando vectores de frecuencia basados en descriptores locales
Ingeniare. Revista Chilena de Ingeniería, vol. 24, núm. 1, enero, 2016, pp. 124-134
Universidad de Tarapacá
Arica, Chile

Clasificación de género utilizando vectores de frecuencia basados en descriptores locales

Gender classification using frequency vectors based on local descriptors

Eduardo Aguilar-Torres[1] Juan Bekios-Calfa[1]

Recibido 24 de noviembre de 2014, Aceptado 11 de mayo de 2015
Received: November 24, 2014 Accepted: May 11, 2015

RESUMEN

La clasificación demográfica, y en particular el reconocimiento de género, es un tema de bastante interés para los investigadores debido a su importancia en diversas aplicaciones, tales como, en áreas de vigilancia, reconocimiento de rostros, indexación de vídeos, estudios de marketing dinámico, entre otras. Éste artículo propone una nueva forma de llevar a cabo la clasificación de género usando vectores de frecuencia basados en descriptores locales SIFT o SURF. El objetivo es poder determinar si los vectores de frecuencia contienen información discriminante. El entrenamiento y la validación de los modelos de clasificación se harán sobre la base de datos Multi-PIE, la cual contiene imágenes de caras tomadas en condiciones de laboratorio, disponibles con cambios de iluminación, pose y expresiones. Para el desarrollo experimental solo se consideran las imágenes capturadas con la iluminación normal de la sala, con los sujetos con expresión neutral y 11 cambios de pose. Los resultados obtenidos validan que los modelos propuestos contienen información discriminante y además mantienen una precisión estable en la clasificación de género sobre imágenes con variaciones de pose. Esto último es sumamente relevante, ya que en condiciones de la vida real difícilmente se van adquirir imágenes de caras frontales, más bien la mayoría tendrán cambios de perspectiva, rotación y cambios de iluminación, por lo tanto se requiere un modelo robusto a estas condiciones.

Palabras clave: Clasificación demográfica, reconocimiento de género, clasificador lineal, descriptor local, vector de frecuencia.

ABSTRACT

The demographic classification, and gender recognition specifically, is a topic of considerable interest to researchers because of its importance in various applications, such as in areas of surveillance, face recognition, indexing videos, dynamic marketing studies, among other. This paper proposes a new way of carry through gender classification using a frequency vector based on local descriptors SIFT or SURF. The goal is to determine if the frequency vectors contain discriminant information. The training and validation of classification models will be based on Multi-PIE data, which contains face images taken under laboratory conditions, available with illumination changes, pose and expressions. For experimental development only considered images captured with normal room lighting, subjects with neutral expression and 11 changes of pose. The results validated that the proposed models contain discriminant information and also maintain a stable accuracy in gender classification on images with pose variations. The latter is extremely important, because in real life conditions are unlikely to acquire images of frontal faces, rather most will change perspective, rotation and illumination changes, therefore a robust model to these conditions is required.

Keywords: Demographic classification, gender recognition, linear classifier, local descriptor, frequency vector.

[1] Departamento de Ingeniería de Sistemas y Computación. Universidad Católica del Norte. 1270709. Antofagasta, Chile. E-mail: eaguilar02@ucn.cl; juan.bekios@ucn.cl

INTRODUCCIÓN

Los atributos visuales definen un conjunto de propiedades observables que dotan de significado a las imágenes, y a partir de éstos podemos entender mejor el entorno físico que nos rodea. En la literatura, las representaciones basadas en atributos han recibido una considerable atención, estos últimos años. Debido a que son utilizadas con éxito en problemas de recuperación de imágenes desde una base de datos [1], en el reconocimiento de objetos [2], en la descripción de objetos desconocidos [3], incluso para aprender nuevos modelos de objetos no entrenados, a través de sus descriptores [3-4]. En el caso particular de los atributos faciales, estos han tenido un papel clave en aplicaciones de interacción hombre-máquina, recuperación de imágenes y video, y vigilancia. También se han utilizado en la verificación de imágenes faciales [5]. Es por esto, que existe un gran interés en encontrar atributos faciales interesantes como lo son el estilo del pelo, la expresión facial, accesorios utilizados, etc.

En el caso, de la estimación de atributos faciales para género, edad y etnia, entre los más importantes, podemos encontrar en la literatura dos aproximaciones: Una **basada en apariencia** [6-7], que utiliza toda la información de una cara recortada desde una imagen, la cual, es normalizada en tamaño e iluminación. Opcionalmente, se puede agregar una máscara para eliminar el efecto del fondo que está detrás de la cara. Por último, se utiliza toda la información resultante como vector de entrada para la clasificación. Otra, la aproximación **basada en características** [6, 8] (*feature-based*) extrae un conjunto discriminante de características de la cara que son utilizadas para la clasificación. La clasificación de atributos faciales, y en particular el reconocimiento de género, estudiada por años [6, 9-14] y SEXNET [12] como el primer intento para reconocer el género de una persona a partir de la imagen de una cara. Dentro de este contexto, una de las publicaciones más relevantes fue la presentada por Moghaddam y Yang [9], quienes propusieron el mejor algoritmo para reconocer género, a esa fecha, a partir de imágenes digitales, en términos de tasas de acierto. Ellos adoptaron una **aproximación basada en apariencia** y utilizaron un clasificador denominado Máquina de Soporte Vectorial (*Support Vector Machine*), con una función de kernel de base radial (*Radial Basis Function Kernel*), SVM+RBF [9], para la clasificación no lineal de datos [15-16]. Los resultados publicados evidencian una tasa de reconocimiento sobre género del 96,6% para la clasificación de 1775 imágenes seleccionadas desde la base de datos FERET [17], donde fueron utilizadas imágenes recortadas de la cara alineadas automáticamente. Se utilizó validación cruzada 5-*fold* para medir la eficiencia del clasificador. Adicionalmente, Baluja y Rowley [10] informaron de un sesgo en el trabajo presentado por Moghaddam y Yang [9] causado por el uso de individuos con la misma identidad en los diferentes *fold* utilizados para la validación cruzada. En el mismo experimento, ellos [10] lograron un 93,5% de tasa de acierto utilizando SVM+RBF con alineación manual y una validación cruzada 5-*fold* donde se considera sujetos diferentes para el entrenamiento y pruebas de cada *fold*. En [7] se establece que esta aproximación tienen un rendimiento del 93,5% aproximadamente sobre FERET, utilizando distintos clasificadores lineales y no lineales. Para aproximaciones **basadas en características**, se utilizan las diferencias de niveles de gris a partir de un par de pixeles [10], Haar-like wavelets [6, 13], bancos de filtros multiescala (multi-scale filter banks) [14] o Locally Binary Patterns (LBP)[6, 11], para reconocer el género a partir de una cara o rostro humano. Shakhmarovich [13] logró un 79% y un 79,2% de precisión en la clasificación de género y origen étnico, respectivamente, sobre un conjunto complicado de imágenes obtenidas desde la web. Asimismo, Baluja y Rowley [10] utilizaron comparaciones de parejas de pixeles en niveles de gris con clasificadores débiles (*weak classifiers*) sobre un esquema de aprendizaje basado en AdaBoost. Ellos utilizaron imágenes alineadas manualmente desde la base de datos de Color FERET, específicamente las galerías "fa" y "fb", donde lograron una precisión del 94%. Sus clasificadores son 50 veces más rápidos que la solución SVM de Moghaddam y Yang [9]. Por su parte, Mäkinen y Raisamo [6] realizaron un conjunto de experimentos utilizando 411 imágenes (304 para entrenamiento y 107 para pruebas) de la base de datos FERET. Ellos compararon las aproximaciones basadas en apariencia y características, con imágenes alineadas y no alineadas, entre los experimentos más importantes. En éstos se obtuvieron resultados de rendimientos similares a las aproximaciones basadas en características (AdaBoost) y las basadas en apariencia (Utilizando clasificadores

SVM+RBF). Ellos reportaron 86% y 82,62% como sus mejores tasas de acierto para imágenes de caras escaladas a un tamaño estándar de 36×36 y 24×24 pixeles respectivamente, utilizando una aproximación basada en apariencia y un clasificador SVM+RBF. En [18] también se comparan ambas aproximaciones, utilizando dos bases de datos, LFW [19] y GROUPS [20], capturadas en condiciones de adquisición no controladas ("*in the wild*") en imágenes de 105×90 y 120×105 donde los mejores resultados obtenidos fueron 79,16% y 86,61% para la aproximación basada en apariencia y características sobre la base de datos GROUPS; 89,24% y 93,83% respectivamente sobre la base de datos LFW. En [21] se utiliza una aproximación basada en características, utilizando la fusión de diferentes clasificadores sobre las bases de datos de adquisición no controladas (LFW, GROUPS y MORPH [22]) con mejoras superiores al 3% con respecto a [18]. En el estudio del reconocimiento de género sobre imágenes con cambios bruscos en la pose de la cara [23] se obtuvo una tasa de acierto entre el 84,31% y el 88,04% sobre la base de datos Multi-PIE [24]. En [25] se obtuvo un 83,7% de tasa de acierto sobre FERET, utilizando un novedoso método basado en características que utiliza zonas específicas de las imágenes descritas con SIFT [26]. Finalmente, de los resultados encontrados en la literatura se obtiene que las aproximaciones **basadas en apariencia** y **características** tienen un comportamiento similar cuando se validan sobre bases de datos que fueron capturadas en condiciones controladas [7, 11]. Sin embargo, cuando la base de datos contiene imágenes capturadas en condiciones no controladas, o con cambios de apariencia bruscos, los rendimientos mejoran cuando se utiliza una **aproximación basada en características** [18, 21, 27].

El siguiente estudio tiene como objetivo probar una aproximación **basada en características** que permita mejorar la robustez de un clasificador de género, utilizando descriptores locales SIFT, SURF y ORB. Nuestra hipótesis de trabajo es que podemos construir vectores de frecuencia, utilizando palabras visuales basadas en SIFT, SURF y ORB que contengan información discriminante y cuyo rendimiento pueda ser comparable a los clasificadores de apariencia en términos de tasa de acierto. Creemos que el principal aporte de ésta es [25]: 1) No se necesita toda la cara para realizar la clasificación, lo que la hace robusta a oclusiones. 2) Los descriptores son independientes de la pose y la escala, lo que permite mejorar la robustez en la clasificación de género cuando las condiciones de apariencia de la cara cambien.

El resto de este trabajo está dividido en 5 partes: La primera, define los descriptores locales SIFT, SURF y ORB. La segunda, define el modelo de espacio vectorial que sirve como base para construir los vectores de frecuencia. La tercera, explica cómo fueron desarrollados los experimentos. La cuarta, muestra los resultados obtenidos. Finalmente, en la última parte se detallan las conclusiones logradas al utilizar esta aproximación y el trabajo futuro a desarrollar.

DESCRIPTOR LOCAL

Los descriptores locales son aquellos que no utilizan toda la imagen para construir un vector de característica, sino, más bien se enfocan en múltiples regiones determinadas o detectadas como puntos de interés. El proceso que se realiza es en primer lugar localizar un punto de interés en una imagen y luego se analiza la vecindad de los pixeles sobre ese punto de interés en una región definida, obteniendo así un conjunto de vectores que describen cada imagen. El hecho de no considerar todos los pixeles de una imagen sino que solo los puntos de interés detectados, permiten contar con un clasificador robusto a oclusiones, cambios de perspectivas y a diversos ruidos que pudieran estar presentes.

A continuación se describe a grandes rasgos el algoritmo de los detectores y descriptores utilizados en la fase experimental.

SIFT (*Scale-invariant feature transform*)

El algoritmo SIFT diseñado por Lowe en el año 2004 [26] es un detector y descriptor que permite a partir de una imagen construir una representación compuesta de puntos de interés invariantes a la escala y rotación.

El algoritmo se compone de cuatro pasos para generar el conjunto de características asociado a la imagen. El primer paso consiste en buscar puntos de interés sobre todas las escalas y posiciones de la imagen representada en diferentes escalas. Esto se implementa eficientemente mediante el uso de una función de diferencia gaussiana para identificar

potenciales puntos que son invariables a escala y orientación. Luego, en cada punto detectado, un modelo detallado es entrenado para determinar aquellos puntos que se mantienen invariantes en cuanto a cambios de escala y localización. Para ello se estudia cada pixel y se realiza una comparación con los pixeles vecinos. Los puntos de interés son seleccionados en base a las medidas de su estabilidad. Lo siguiente es asignar una o más orientaciones, para cada punto de interés, de acuerdo a las direcciones del gradiente y a la zona que rodea dicho punto. Finalmente, se calcula un descriptor para cada punto de interés de la imagen local. Los gradientes de la imagen son medidos para seleccionar la escala y región alrededor del punto. Estos son transformados en una representación que permite ser invariable a significativos cambios de iluminación y perspectiva.

Por cada punto de interés detectado se toma una vecindad de 16x16 pixeles alrededor del punto. La región circular alrededor del punto de interés es dividido en sub-regiones de 4x4 pixeles sin sobreponerse, donde se calculan los histogramas de las orientaciones del gradiente. Para cada sub-región se crea un histograma de 8 bin. Esto se traduce en un vector de características de 128 dimensiones (4x4x8) para cada punto de interés.

SURF (*Speeded-Up Robust Features*)

El algoritmo SURF [28] fue diseñado en el año 2006 como una versión acelerada de SIFT. Al igual que SIFT, es un detector y descriptor que permite a partir de una imagen construir una representación compuesta de puntos de interés invariantes a la escala y rotación, pero no es tan bueno con imágenes que presentan cambios de perspectiva y cambios de iluminación.

El algoritmo de SURF contempla tres pasos. El primer paso consiste en detectar puntos de interés por medio de un *box filter* que, al igual que en el caso de SIFT con la diferencia gaussiana, aproxima al filtro *Laplacian of Gaussian* para identificar potenciales puntos que son invariables a los cambios de escala. Debido al uso de *box filter* y las imágenes integrales, no se tiene que aplicar iterativamente el mismo filtro a la salida de una capa previamente filtrada, sino que puede aplicar tales filtros de cualquier tamaño exactamente a la misma velocidad directamente en la imagen original. Por lo tanto, el espacio escalar se analiza por el aumento del tamaño del filtro en lugar de reducir la imagen original. Luego de detectar los puntos de interés, para que estos sean invariantes a la rotación, primero se calcula las respuestas de *Haar-wavelet* en las direcciones x e y, en un radio de 6s alrededor del punto. El parámetro s es la escala con que el punto de interés fue detectado. Luego se calcula y pondera las respuestas wavelet con una función Gaussiana centrada en el punto de interés, la respuesta se representa como vector. La orientación dominante se determina mediante la suma de todas las respuestas dentro de una ventana deslizante que cubre la orientación de un ángulo de 60°. Se suman las respuestas horizontales y verticales dentro de la ventana, formando nuevos vectores, donde el más largo de estos se le extrae su orientación para el punto de interés. Finalmente se extraen los descriptores para cada punto de interés, donde en primer lugar se construye una ventana centrada en el punto de interés, y orientada a lo largo de la dirección seleccionada por la suma de todas las respuestas horizontales y verticales dentro de la ventana. La región se divide regularmente en pequeñas sub-regiones de 4x4. Para cada sub-región las respuestas wavelet horizontales y verticales son tomadas para construir el vector de característica. Este vector queda representado con 64 dimensiones pero puede ser extendido a 128 dimensiones lo que permite obtener características más distintivas pero se disminuye la rapidez de la extracción.

ORB (Oriented FAST and Rotated BRIEF)

Oriented FAST and Rotated BRIEF (ORB) es un detector y descriptor, invariante a la rotación y resistente al ruido, desarrollado por [29]. ORB surge como una alternativa eficiente a SIFT y SURF ya que dentro de sus principales características se destaca su bajo costo de computación, mantiene un buen rendimiento y, a diferencia de SIFT y SURF, no se debe pagar por su uso.

ORB se compone por la fusión de un detector de puntos de interés y un descriptor, denominados FAST [30] y BRIEF [31] respectivamente. A estos algoritmos se les incorporan una serie de modificaciones para mejorar el rendimiento y lograr un algoritmo invariante a la rotación y robusto a los cambio de escala.

FAST se utiliza para encontrar los puntos de interés, de los cuales por medio de la medida de esquina de Harris se seleccionan los N mejores. Para incorporar

la invariancia a la escala al descriptor, se emplea una pirámide de escala de la imagen para generar puntos de interés para cada nivel de la pirámide. También se computa la orientación por medio del centroide de la intensidad. Respecto al descriptor, ORB utiliza una versión dirigida de BRIEF de acuerdo a la orientación de los puntos de interés dejándolo invariante a la rotación.

VECTOR DE FRECUENCIA

Inspirados en los modelos de recuperación de información y, específicamente, en el área de minería de textos (*text mining*), hemos utilizado y adaptado el **modelo de espacio vectorial** [32] que es utilizado para la representación y búsqueda de documentos. El modelo representa los documentos de texto como vectores definidos como histogramas de frecuencias de palabras que se encuentran en el texto. En otras palabras, construimos una función *f*: $T \rightarrow N$ donde T es un conjunto de términos presentes en un documento, y N el vector que representa la importancia de cada término. Si cada término representa una dimensión en un espacio vectorial. Las dimensiones del vector resultante representan la importancia de cada término en el documento y que son representados por un escalar. Este modelo también es llamado representación *bag-of-word*, donde el orden y la localización de las palabras es ignorado [33].

Para poder aplicar las técnicas de recuperación información a imágenes fue necesario crear un elemento visual similar a las palabras utilizadas en los documentos. Esto se logra utilizando SIFT, SURF y ORB para la detección y descripción de puntos relevantes por imagen. Los descriptores de un conjunto de imágenes son almacenados para construir un diccionario de **palabras visuales**. El diccionario se construye identificando grupos de descriptores por medio de un algoritmo de *clustering* tal como k-medias. El algoritmo asigna un número de *cluster* o de pertenencia a los nuevos descriptores. Finalmente, nuestro modelo de espacio vectorial se construye a partir del modelo definido por k-medias y que actúa como un diccionario de palabras. El vector resultante tiene tantas dimensiones como *clusters* sean definidos en el modelo k-medias. En nuestros experimentos el vector se construye calculando la frecuencia con la que aparecen las palabras visuales en la imagen a partir de los descriptores obtenidos por los algoritmos SIFT, SURF u ORB, Figura 1.

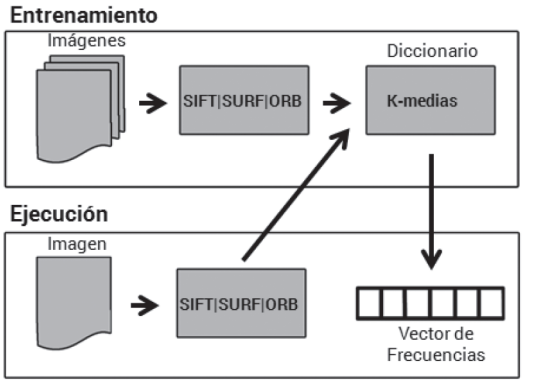

Figura 1. Construcción de vector de frecuencias utilizando descriptores SIFT, SURF u ORB y k-medias.

EXPERIMENTO

En esta sección se muestra el procedimiento realizado para la clasificación de género utilizando modelos basados en apariencia y modelos basados en características locales.

Base de datos

El conjunto de datos es de vital importancia para poder construir el modelo de clasificación, por medio del entrenamiento y validación del clasificador. Los datos utilizados corresponden a imágenes faciales de laboratorio adquiridas en condiciones controladas disponibles en la base de datos Multi-PIE [24] (Múltiples poses, iluminaciones y expresiones).

La base de datos Multi-PIE, creada en el año 2009 por investigadores de *Carnegie Mellon University*, contiene imágenes de caras adquiridas en diferentes condiciones de iluminación, pose y expresiones faciales. En total almacena 755370 imágenes, las cuales se tomaron de 337 individuos, 235 hombres y 102 mujeres, en cuatro sesiones en un período de 6 meses. Cabe destacar que los individuos son de diferentes etnias (60% europeos-americanos, 35% asiáticos, 3% africanos y 2% otros) y su edad promedio es de 27,9 años. Específicamente las imágenes con variación de pose e iluminación se capturaron con 15 cámaras en forma simultánea aplicando 19 cambios de iluminación, 13 colocadas a la altura de la cabeza del sujeto, con una variación de 15% grados entre ellas, y 2 sobre el sujeto tal

como se muestra en la Figura 2. La mayoría de las cámaras fueron producidas por Sony (11 de 15) y el resto por Panasonic (posiciones: 11_0, 08_1, 19_1 y 24_0).

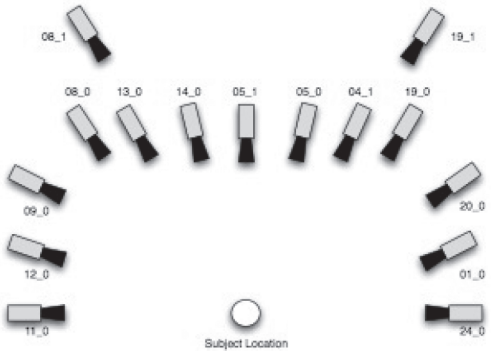

Figura 2. Etiquetas y distribución aproximada de las cámaras dentro de la sala.

En el experimento se utilizan todas las imágenes con variaciones de pose, iluminación normal de la sala y expresión facial neutral, adquiridas por las cámaras Sony en la primera sesión. Sobre el conjunto original de imágenes, que tienen una resolución de 640x480 pixeles, se les aplica el algoritmo de Viola & Jones [34] con la finalidad de detectar la cara y utilizar solo esa región. Se obtienen en total 2625 imágenes, 1865 hombres y 760 mujeres, con una resolución media de 200x200 pixeles (Ver Tabla 1). Un ejemplo de imágenes resultantes con variaciones de poses se puede apreciar en la Figura 3.

Tabla 1. Frecuencia de imágenes de caras de hombres y mujeres distribuidas por pose.

Pose	Hombre	Mujer
12_0	164	51
09_0	173	69
08_0	173	75
13_0	166	74
14_0	173	76
05_1	173	76
05_0	173	76
04_1	161	73
19_0	173	73
20_0	172	68
01_0	164	49

Figura 3. Imágenes etiquetadas adquiridas por 11 cámaras simultáneamente, detectadas con Viola & Jones y pre-procesadas con ecualización del histograma.

Modelo basado en apariencia

Los modelos basados en apariencia usan un enfoque holístico para describir una imagen, es decir, utilizan todos los pixeles de la imagen para generalizar el objeto que se desea reconocer en un simple vector.

El modelo basado en apariencia utilizado es bastante simple y considera técnicas comunes de pre-procesamiento de imágenes. A continuación se muestra paso a paso el proceso realizado:

1. Se ecualiza el histograma de las imágenes detectadas por Viola & Jones para mejorar su contraste.
2. Se redimensionan las imágenes a un tamaño de 25x25 pixeles.
3. Se construye un vector de característica de 625 dimensiones, una dimensión por pixel, por cada imagen.

Modelo basado en características locales

A diferencia de los modelos basados en apariencia, un modelo basado en características locales no utiliza todos los pixeles de la imagen, sino que se basan en la extracción de características de múltiples regiones de la imagen detectadas como puntos de interés. En términos generales, cada imagen se describe por medio de un conjunto de vectores de características asociados a cada punto de interés detectado.

Se construyen tres modelos basados en características locales a partir de los detectores y descriptores SIFT, SURF y ORB. Para estos se utiliza el mismo proceso, el cual se muestra a continuación:

1. Se detectan los puntos característicos de las imágenes obtenidas con Viola & Jones, sin redimensionar.
2. Se obtienen los descriptores asociados a cada punto detectado en las imágenes.

3. Se toma todos los descriptores y se agrupan en 1000 *cluster* por medio del algoritmo de k-medias.
4. Finalmente por cada imagen se construye un vector de característica de 1000 dimensiones que representa la frecuencia de los *cluster* asociados a cada punto de interés.

Los detectores y descriptores SIFT, SURF y ORB se utilizan desde OpenCV, configurando los parámetros inicialmente con lo recomendado en sus publicaciones. Lamentablemente SIFT fue diseñado para la detección y reconocimiento de objetos en general, es decir objetos rígidos y con transiciones bruscas entre sus diferentes lados. Sin embargo las caras no son rígidas ni presentan transiciones bruscas por lo que utilizar SIFT con las configuraciones propuestas no es recomendable [35]. Por lo tanto se configura SIFT basado en el enfoque propuesto por *Geng & Jiang* [36] denominado *Volume-SIFT* para eliminar los puntos no confiables, los cuales se identifican analizando la escala de los puntos de interés a diferencia del algoritmo original que lo hace por medio del contraste. Por otro lado, en el caso de SURF se utilizan los parámetros recomendados por su autor, variando solo el umbral del *Hessiano* de manera tal de obtener una cantidad de descriptores similares que con SIFT. De igual manera con ORB, se utilizan los parámetros recomendados por su autor limitando la cantidad de descriptores a la misma que con SIFT.

En la Figura 4 se pueden ver los puntos de interés detectados con SIFT, SURF y ORB en la primera, segunda y tercera fila respectivamente.

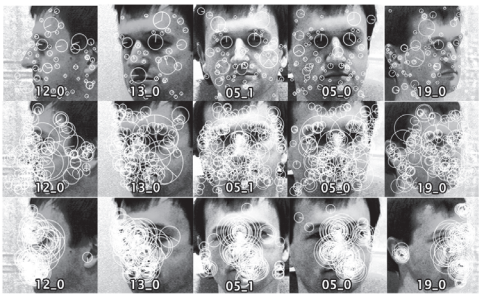

Figura 4. Puntos de interés detectados con SIFT, SURF y ORB, sobre imágenes de un mismo sujeto con variación de pose.

Entrenamiento y validación

En ambos modelos de clasificación se utiliza la estrategia de entrenamiento y validación denominado LOPO (*Leaft One Person Out*). Es decir, el conjunto de datos se divide en p secciones, donde p corresponde al número de personas que tiene la base de datos, en las que 1 sección se utiliza para validar y las p-1 restantes para entrenar. El proceso se realiza p veces cambiando la sección de validación.

En el caso del modelo de apariencia, en cada iteración se toman los vectores de características asociados a las secciones de entrenamientos y a estos se le aplica la reducción de dimensión por medio de PCA + LDA, descrito en [7], y luego se entrena el clasificador lineal *Gaussian Naive Bayes*. Después de haber entrenado el clasificador, se utilizan los datos de la sección de prueba para validar y por consiguiente se compara la clase estimada con la real para determinar la precisión. El número de componentes utilizados en PCA se determinan empíricamente seleccionando el que mejora la precisión media obtenida por las iteraciones.

En el caso del modelo basado en características locales el proceso es similar al anterior, con la salvedad de que en cada iteración se tienen que realizar los pasos 3 y 4 mencionados en la sección previa. Es decir por cada iteración se tienen que tomar los descriptores asociados a las imágenes de entrenamiento con los que se generan 1000 *cluster* y luego construir los vectores de características asociado a cada imagen. Luego la fase de entrenamiento y validación del clasificador se mantiene igual a la descrita previamente.

La precisión obtenida en la clasificación de género utilizando modelos basados en apariencia y modelos basados en características, se pueden ver gráficamente en la Figura 6 y numéricamente en la Tabla 3.

RESULTADOS

En esta sección se muestran los resultados obtenidos en la fase experimental.

En la Tabla 2 se observa el número de componentes principales utilizados para reducir la dimensión del vector de características antes de realizar la clasificación, y las tasas de aciertos globales obtenidas en cada uno de los experimentos. Se destaca con

negrilla el algoritmo con el cual se obtuvo el peor rendimiento en la clasificación.

Tabla 2. Tasa de acierto globales obtenidas en la clasificación de género utilizando modelos basados en apariencia y modelos basados en características.

	PCA	Tasa de acierto
Apariencia	40	85,34% ± 26,64%
SIFT	210	84,67% ± 24,08%
SURF	140	85,10% ± 24,30%
ORB	130	**78,09% ± 28,86%**

La Figura 5 representa la relación de una pose específica de Multi-PIE con su tasa de acierto en la predicción de género. En el eje de la abscisa se ordenan las poses desde el perfil izquierdo variando la rotación de la cara hasta el perfil derecho.

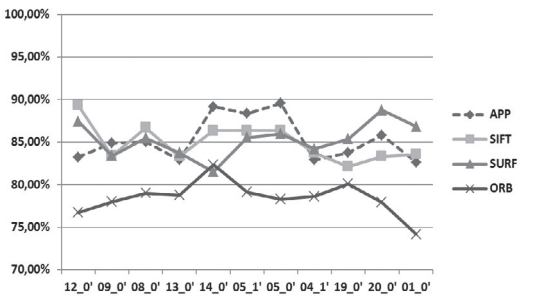

Figura 5. Gráfico que muestra la precisión global obtenida por los cuatro modelos de clasificación propuestos, distribuidos por pose.

Las tasas de acierto junto con su desviación estándar para cada pose se pueden revisar en la Tabla 3. Además, se destacan los máximos rendimientos locales, es decir por pose, y el rendimiento general. Cabe destacar, que no se muestra con más detalle los resultados obtenidos con ORB dado a la baja precisión presentada al utilizar este descriptor de características.

Las Tablas 4, 5 y 6 corresponden a los resultados estratificados por pose y género para los modelos basados en apariencia, en características locales con SIFT y en características locales con SURF respectivamente. En las tablas se destacan los rendimientos mínimos y máximos por género para cada modelo de clasificación. También se destacan los rendimientos generales.

Tabla 3. Tasa de acierto obtenida en la clasificación de género utilizando modelos basados en apariencia y modelos basados en características separados por la pose de los individuos.

Pose	Apariencia (%)	SIFT (%)	SURF (%)
12_0'	83,26 ± 37,42	**89,30 ± 30,98**	87,44 ± 33,22
09_0'	**84,85 ± 35,78**	83,40 ± 37,28	83,40 ± 37,28
08_0'	85,08 ± 35,7	**86,69 ± 34,03**	85,48 ± 35,30
13_0'	82,92 ± 37,71	83,33 ± 37,35	**83,75 ± 36,97**
14_0'	**89,16 ± 31,15**	86,35 ± 34,41	81,53 ± 38,89
05_1'	**88,35 ± 32,14**	86,35 ± 35,24	85,54 ± 35,24
05_0'	**89,56 ± 30,64**	86,35 ± 34,41	85,94 ± 34,41
04_1'	82,91 ± 37,73	83,76 ± 36,96	**84,19 ± 36,56**
19_0'	83,74 ± 36,98	82,11 ± 38,4	**85,37 ± 35,42**
20_0'	85,83 ± 34,94	83,33 ± 37,35	**88,75 ± 31,66**
01_0'	82,63 ± 37,98	83,57 ± 37,14	**86,85 ± 33,87**

Tabla 4. Precisión obtenida en la clasificación de género utilizando modelos basados en apariencia separados por pose y género del individuo.

	Apariencia	
Pose	**Hombre**	**Mujer**
12_0'	92,68% ± 26,12%	52,94% ± 50,41%
09_0'	**98,84% ± 10,72%**	49,26% ± 49,99%
08_0'	95,95% ± 19,76%	60,00% ± 49,32%
13_0'	98,19% ± 13,36%	48,65% ± 50,32%
14_0'	94,22% ± 23,40%	77,63% ± 41,95%
05_1'	**90,75% ± 29,06%**	**82,89% ± 37,91%**
05_0'	93,64% ± 24,47%	80,26% ± 40,07%
04_1'	98,14% ± 13,56%	49,32% ± 50,34%
19_0'	94,22% ± 23,40%	58,90% ± 49,54%
20_0'	97,67% ± 15,12%	55,88% ± 50,02%
01_0'	93,29% ± 25,09%	**46,94% ± 50,42%**
General	**95,06% ± 11,05%**	**63,21% ± 36,79%**

Tabla 5. Precisión obtenida en la clasificación de género utilizando modelos basados en características locales con SIFT separados por pose y género del individuo.

	SIFT	
Pose	**Hombre**	**Mujer**
12_0'	92,07% ± 27,10%	**80,39% ± 40,10%**
09_0'	97,11% ± 16,80%	48,53% ± 50,35%
08_0'	**97,69% ± 15,07%**	61,33% ± 49,03%
13_0'	95,18% ± 21,48%	56,76% ± 49,88%
14_0'	95,95% ± 19,76%	64,47% ± 48,18%
05_1'	**89,02% ± 31,36%**	77,63% ± 41,95%
05_0'	93,64% ± 24,47%	69,74% ± 46,24%
04_1'	95,03% ± 21,80%	58,90% ± 49,54%
19_0'	**97,69% ± 15,07%**	**45,21% ± 50,11%**
20_0'	95,93% ± 19,82%	51,47% ± 50,35%
01_0'	92,68% ± 26,12%	53,06% ± 50,42%
General	**94,77% ± 7,98%**	**61,68% ± 31,62%**

Tabla 6. Precisión obtenida en la clasificación de género utilizando modelos basados en características locales con SURF separados por pose y género del individuo.

	SURF	
Pose	**Hombre**	**Mujer**
12_0'	97,56% ± 15,47%	54,90% ± 50,25%
09_0'	97,69% ± 15,07%	**47,06% ± 50,28%**
08_0'	95,95% ± 19,76%	61,33% ± 49,03%
13_0'	95,78% ± 20,16%	56,76% ± 49,88%
14_0'	**88,44% ± 32,07%**	65,79% ± 47,76%
05_1'	93,64% ± 24,47%	67,11% ± 47,30%
05_0'	95,38% ± 21,06%	64,47% ± 48,18%
04_1'	89,44% ± 30,83%	**72,60% ± 44,91%**
19_0'	98,27% ± 13,09%	54,79% ± 50,11%
20_0'	**100,00% ± 0,00%**	60,29% ± 49,29%
01_0'	94,51% ± 22,84%	61,22% ± 49,23%
General	**95,16% ± 8,73%**	**62,20% ± 31,83%**

CONCLUSIONES

En este artículo, se demuestra empíricamente que el modelo de clasificación basado en vectores de frecuencia, construidos a partir de características locales, por medio de SIFT y SURF, entrega información discriminante y obtiene resultados similares a los modelos basados en apariencia (ver Tabla 2). No se puede decir lo mismo en el caso de utilizar ORB dado al bajo desempeño mostrado con relación a la tasa de acierto, logrando con este algoritmo alrededor de un 7% menos con relación a los otros tres modelos de clasificación.

Al analizar los resultados de las Tablas 4, 5 y 6, se puede ver que los modelos de clasificación propuestos entregan mejor precisión al clasificar hombres que al clasificar mujeres. Claramente esto ocurre al utilizar una base de datos desbalanceada en género, lo que hace que los clasificadores tiendan a ajustar el modelo a la clase dominante para mejorar la precisión.

De los datos obtenidos en la Tabla 3, se confirma que los modelos basados en apariencia predicen mejor sobre caras en posiciones frontales (14_0, 05_1, 05_0), a diferencia de los modelos basados en características propuestos, los cuales mantienen un resultado regular con las variaciones de pose. Esto también es posible de ver en la Figura 5, en la que se muestra que los modelos basados en características propuestos presentan una mejor precisión que los basados en apariencia en cambios bruscos de pose, y los de apariencia una mejor precisión en caras frontales.

A futuro queda pendiente evaluar los modelos basado en características propuestos sobre imágenes tomadas en entornos de la vida real, entornos en que los modelos basados en apariencia no funcionan del todo bien, y buscar una mejor configuración de los parámetros SIFT, SURF y el número de *cluster* a utilizar.

REFERENCIAS

[1] F.X. Yu, R. Ji, M.H. Tsai, G. Ye and S. F. Chang. "Weak attributes for large-scale image retrieval". In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2949-2956. 2012.

[2] K. Duan, D. Parikh, D. Crandall and K. Grauman. "Discovering localized attributes for fine-grained recognition". In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3474-3481. 2012.

[3] A. Farhadi, I. Endres, D. Hoiem and D. Forsyth. "Describing objects by their attributes". In 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009, pp. 1778-1785. 2009.

[4] C.H. Lampert, H. Nickisch and S. Harmeling. "Learning to detect unseen object classes by between-class attribute transfer". In 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. CVPR Workshops 2009, pp. 951-958. 2009.

[5] N. Kumar, A.C. Berg, P.N. Belhumeur and S.K. Nayar. "Attribute and simile classifiers for face verification". In Proceedings of the IEEE International Conference on Computer Vision, pp. 365-372. 2009.

[6] E. Mäkinen and R. Raisamo. "Evaluation of gender classification methods with automatically detected and aligned faces". IEEE Trans. Pattern Anal. Mach. Intell. Vol. 30 Nº 3, pp. 541-7. March, 2008.

[7] J. Bekios-Calfa, J.M. Buenaposada and L. Baumela. "Revisiting linear discriminant techniques in gender recognition". IEEE Trans. Pattern Anal. Mach. Intell. Vol. 33 Nº 4, pp. 858-64. May, 2011.

[8] Z. Yang and H. Ai. "Demographic classification with local binary patterns". Adv. Biometrics. 2007.

[9] B. Moghaddam and M.-H. Yang. "Learning gender with support faces". Pattern Anal. Mach. Intell. IEEE Trans. Vol. 24 Nº 5, pp. 707-711. May, 2002.

[10] S. Baluja and H.A. Rowley. "Boosting Sex Identification Performance". Int. J. Comput. Vis. Vol. 71 Nº 1, pp. 111-119. June, 2006.

[11] E. Mäkinen and R. Raisamo. "An experimental comparison of gender classification methods". Pattern Recognit. Lett. Vol. 29 Nº 10, pp. 1544-1556. July, 2008.

[12] B.A. Golomb, D.T. Lawrence and T.J. Sejnowski. "Sexnet: A Neural Network Identifies Sex from Human Faces". In Advances in neural information processing systems 3, pp. 572-577. 1990.

[13] G. Shakhnarovich, P.A. Viola and B. Moghaddam. "A unified learning framework for real time face detection and classification". Proc. Fifth IEEE Int. Conf. Autom. Face Gesture Recognit, pp. 16-23. 2002.

[14] A. Lapedriza, M.J. Maryn-Jimenez and J. Vitria. "Gender Recognition in Non Controlled Environments". In 18th International Conference on Pattern Recognition (ICPR'06), pp. 834-837. 2006.

[15] B. Liu. "Web data mining exploring hyperlinks, contents, and usage data". Second edition. Springer. Berlin, Germany, pp. 109-120. 2011.

[16] C.M. Bishop. "Pattern Recognition and Machine Learning". Springer, pp. 325-345. 2006.

[17] P.J. Phillips, S.A. Rizvi and P.J. Rauss. "The FERET evaluation methodology for face-recognition algorithms". IEEE Trans. Pattern Anal. Mach. Intell. Vol. 22 Nº 10, pp. 1090-1104. 2000.

[18] P. Dago-Casas, D. Gonzalez-Jimenez and J.L. Alba-Castro. "Single- and cross- database benchmarks for gender classification under unconstrained settings". In 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2152-2159. 2011.

[19] G.B. Huang, M. Ramesh, T. Berg and E. Learned-miller. "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments". Amherst. 2007.

[20] A.C. Gallagher and T. Chen. "Understanding images of groups of people". In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 256-263. 2009.

[21] J. Castrillón-Santana, Modesto Lorenzo-Navarro and E. Ramón-Balmaseda. "Improving Gender Classification Accuracy in the Wild". in Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, CIARP 2013. Vol. 8259. J. Ruiz-Shulcloper and G. Sanniti di Baja, Eds. Springer. Berlin, Germany, pp. 270-277. 2013.

[22] K. Ricanek and T. Tesafaye. "MORPH: A Longitudinal Image Database of Normal Adult Age-Progression". In 7th International Conference on Automatic Face and Gesture Recognition (FGR06), pp. 341-345. 2006.

[23] J. Bekios-Calfa, J.M. Buenaposada and L. Baumela. "Robust gender recognition by

exploiting facial attributes dependencies". Pattern Recognit. Lett. Vol. 36, pp. 228-234. June, 2014.

[24] R. Gross, I. Matthews, J. Cohn, T. Kanade and S. Baker. "Multi-PIE.," Proc. Int. Conf. Autom. Face Gesture Recognit. Vol. 28 Nº 5, pp. 807-813. May, 2010.

[25] M. Toews and T. Arbel. "Detection, localization and sex classification of faces from arbitrary viewpoints and under occlusion". IEEE Trans. Pattern Anal. Mach. Intell. Vol. 31 Nº 9, pp. 1567-81. September, 2009.

[26] D.G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". Int. J. Comput. Vis. Vol. 60 Nº 2, pp. 91-110. November, 2004.

[27] E. Ramón-Balmaseda, J. Lorenzo-navarro and M. Castrillón-Santana. "Gender Classification in Large Databases". In Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Vol. 7441. L. Alvarez, M. Mejail, L. Gomez and J. Jacobo. Eds. Springer. Berlin, Germany, pp. 74-81. 2012.

[28] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool. "Speeded-Up Robust Features (SURF)". Comput. Vis. Image Underst. Vol. 110 Nº 3, pp. 346-359. June, 2008.

[29] E. Rublee, V. Rabaud, K. Konolige and G. Bradski. "ORB: An efficient alternative to SIFT or SURF". In Proceedings of the IEEE International Conference on Computer Vision, pp. 2564-2571. 2011.

[30] E. Rosten and T. Drummond. "Machine learning for high-speed corner detection". In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 3951 LNCS, pp. 430-443. 2006.

[31] M. Calonder, V. Lepetit, C. Strecha and P. Fua. "BRIEF: Binary robust independent elementary features". In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 6314 LNCS, pp. 778-792. 2010.

[32] G. Salton, A. Wong and C. S. Yang. "AVector Space Model for Automatic Indexing". Vol. 18 Nº 11. 1975.

[33] Jan Erik Solem, Programming Computer Vision with Python. O'Reilly Media, p. 264. 2012.

[34] P. Viola and M.J. Jones. "Robust Real-Time Face Detection". Int. J. Comput. Vis. Vol. 57 Nº 2, pp. 137-154. May, 2004.

[35] D.S. Al-Azzawy and S. Al-Azzawy. "Eigenface and SIFT For Gender Classification". J. Wassit Sci. Med. Vol. 5 Nº 1, pp. 60-76. 2012.

[36] C. Geng and X. Jiang. "Face recognition using sift features". In Proceedings - International Conference on Image Processing. ICIP, pp. 3313-3316. 2009.